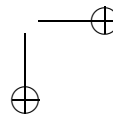
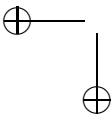
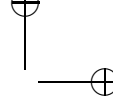
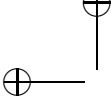


Cartas de control T^2 multivariadas usando R y SAS



Sergio Yáñez Canal, M.Sc.

Nelfi González Álvarez, Ph.D.

José Alberto Vargas Navas, Ph.D.

Cartas de control T^2 multivariadas usando R y SAS



UNIVERSIDAD **NACIONAL** DE COLOMBIA

SEDE MEDELLÍN
FACULTAD DE CIENCIAS
ESCUELA DE ESTADÍSTICA

SEDE BOGOTÁ
FACULTAD DE CIENCIAS
DEPARTAMENTO DE ESTADÍSTICA

Bogotá, D.C. diciembre de 2011

- © Universidad Nacional de Colombia, sede Medellín
Escuela de Estadística
- © Sergio Yáñez Canal
- © Nelfi González Álvarez

- © Universidad Nacional de Colombia, sede Bogotá
Departamento de Estadística
- © José Alberto Vargas Navas

ISBN 978-958-761-055-0

Primera edición, 2011

PREPARACIÓN EDITORIAL E IMPRESIÓN:

Editorial Universidad Nacional de Colombia

www.editorial.unal.edu.co

direditorial@unal.edu.co

Bogotá, Colombia

Prohibida la reproducción total o parcial por cualquier medio
sin la autorización escrita del titular de los derechos patrimoniales

Impreso y hecho en Bogotá, D. C. Colombia

Catalogación en la publicación Universidad Nacional de Colombia

Yáñez Canal, Sergio, 1951-

Cartas de control T^2 multivariadas usando R y SAS / Sergio Yáñez Canal,
Nelfi González Álvarez, José Alberto Vargas Navas. - Medellín: Universidad Na-
cional de Colombia. Facultad de Ciencias. Escuela de Estadística; Bogotá: Univer-
sidad Nacional de Colombia. Facultad de Ciencias. Departamento de Estadística,
2011

xviii, 204 p. : il.

Incluye referencias bibliográficas

ISBN : 978-958-761-055-0

1. Control de calidad - Métodos estadísticos 2. Cartas de control 3. Estadística
industrial 4. Análisis multivariante I. González Álvarez, Nelfi Gertrudis, 1968- II.
Vargas Navas, José Alberto, 1956- III. Tít.

CDD-21 658.562015195 / 2011

Contenido

Prefacio	1
1 Introducción al control estadístico de procesos	5
1.1 Control estadístico de procesos (SPC)	5
1.2 Bosquejo histórico	5
1.3 Algunas precisiones sobre el SPC	6
1.3.1 Control	7
1.3.2 El proceso	7
1.3.3 Estadística	9
2 Cartas de control para mediciones con subgrupos	11
2.1 Principios básicos de las cartas de control	11
2.2 Tipos de errores (riesgos del muestreo)	13

2.3	Aspectos estadísticos básicos de las cartas de control . . .	15
2.3.1	Cartas de control en tiempo real y sobre datos históricos	16
2.4	Construcción de cartas de control	18
2.4.1	Carta R	20
2.4.2	Carta R con límites probabilísticos	20
2.4.3	Carta s	21
2.4.4	Carta s con límites probabilísticos	21
2.4.5	Carta s^2	22
2.4.6	Carta $\bar{X}$	23
2.5	Análisis en SAS usando PROC SHEWHART	24
2.6	Longitud promedio de corridas	41
3	Cartas de control para observaciones individuales	45
3.1	Cartas para observaciones individuales	45
3.1.1	Límites de control para la carta $\bar{X}$	46
3.1.2	Supuestos de la carta $\bar{X}$	48
3.1.3	Ejemplo ilustrativo usando SAS	48
3.2	Cartas de medias móviles	59
3.2.1	Continuación del ejemplo anterior	60
4	Cartas de control multivariadas	63
4.1	¿Por qué control multivariado?	63
4.1.1	Procesos univariados <i>versus</i> procesos multivariados	64

CONTENIDO

IX

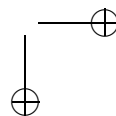
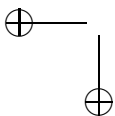
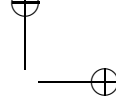
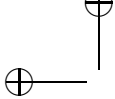
4.1.2	Características deseables de un procedimiento de control multivariado	66
4.2	Aspectos preliminares: estudio del T^2 de Hotelling	68
4.2.1	Variables y observaciones	68
4.2.2	Matriz de datos y estadísticos muestrales multivariados	69
4.2.3	Medidas de dispersión multivariadas	72
4.2.4	Combinaciones lineales	73
4.2.5	Transformaciones	74
4.2.6	Visión geométrica de la matriz de datos	76
4.3	La distribución normal multivariada	77
4.3.1	Densidad normal multivariada	78
4.3.2	Distribución normal bivariada	79
4.3.3	Contornos de densidad constante	79
4.3.4	Algunas propiedades de la distribución normal multivariada	83
4.4	Distancia estadística <i>versus</i> distancia euclidiana	84
4.4.1	Distancia euclidiana	84
4.4.2	Distancia estadística	85
4.5	Estadístico T^2 de Hotelling	87
4.5.1	Algunas propiedades	88
4.5.2	Otros resultados importantes	88

4.6	Evaluación del supuesto de normalidad multivariada	89
4.6.1	Procedimiento 1	90
4.6.2	Procedimiento 2	93
4.6.3	Otro procedimiento: gráfico Q-Q chi cuadrado . . .	99
4.7	El estadístico de control T^2 bajo normalidad	103
4.7.1	Propiedades distribucionales del estadístico T^2 de Hotelling y determinación del límite de control superior (UCL)	103
4.8	Chequeo de supuestos para el uso del estadístico T^2	106
4.8.1	Normalidad multivariada	106
4.8.2	Transformaciones y aproximaciones no paramétricas	108
4.8.3	Tamaños de muestra	117
4.9	Construcción de la carta de control T^2	117
4.9.1	Programación en R	117
4.9.2	Construcción de la base de datos históricos o HDS	126
4.9.3	Procedimientos de recolección de datos	128
4.9.4	Datos faltantes	128
4.9.5	Detección de colinealidad	134
4.9.6	Diagnóstico de no independencia entre observaciones	135
4.10	Fase I de control	142
4.10.1	Depuración bajo normalidad	143

CONTENIDO

XI

4.10.2	Depuración bajo no normalidad	144
4.11	Fase II de control	146
4.11.1	Escogencia de la tasa de falsa alarma	147
4.11.2	Reacción a las señales	148
4.11.3	Interpretación de patrones en la carta T^2	148
4.12	Control mediante componentes principales	149
4.13	Interpretación de señales	160
4.13.1	Descomposición MYT, caso bivariado	161
4.13.2	Descomposición MYT, caso general	166
4.13.3	Propiedades de la descomposición MYT	168
4.13.4	La regresión lineal como medio para mejorar la interpretación de señales atribuibles a términos condicionales de la descomposición MYT	175
5	Métodos robustos para el vector de medias	177
5.1	Introducción	177
5.2	Estimación de los parámetros y algunos métodos	178
5.3	Carta de control T^2 basada en estimadores DG	181
5.3.1	Estimadores DG (Donoho-Gasko)	182
	Bibliografía	188
	Apéndice A Una breve introducción a R	195



Lista de figuras

1.1	Búsqueda de una estrategia para un resultado deseado. . .	8
2.1	Carta de control típica.	12
2.2	Objetivo de una carta de control.	13
2.3	Distribución normal.	15
3.1	Gráfico de probabilidad normal para los datos de la tabla 3.1.	49
3.2	Carta $\bar{X}$ y de rangos móviles para observaciones individuales. σ estimado como $\overline{MR}/d_2$	52
3.3	Carta X-MR para datos de la tabla 3.1, sin incluir las observaciones 11 y 33.	53
3.4	Carta X-MR para datos de la tabla 3.1, sin incluir las observaciones 11, 33 y 36.	55
3.5	Carta $\bar{X}$ y de rangos móviles para observaciones individuales. σ estimado como s/c_4	58

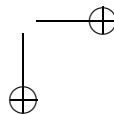
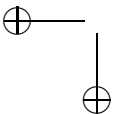
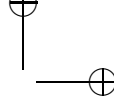
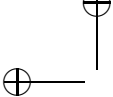
3.6	Carta de control de medias móviles.	62
4.1	Región de control bivariada <i>vs.</i> región de control con dos cartas univariadas.	65
4.2	Carta de control T^2 en la Fase I, con 30 observaciones históricas de un proceso bivariado.	67
4.3	Distribuciones normales bivariadas, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ y $\sigma_{12} = \rho$	80
4.4	Contornos de probabilidad del 30 %, 50 %, 70 % y 90 %, de una normal bivariada con $\boldsymbol{\mu} = (1, 2)^t$, $\sigma_1^2 = 4$, $\sigma_2^2 = 1$ y $\rho_{12} = 0,95$	82
4.5	Representación geométrica de la distancia euclidiana. . . .	85
4.6	Representación geométrica de la distancia estadística de dos variables con distribución normal bivariada, coeficiente de correlación positivo.	87
4.7	Matriz de dispersión con histogramas. Se pueden evaluar las características univariadas y las relaciones por pares entre variables.	92
4.8	Matriz de dispersión con <i>boxplots</i> . Se pueden evaluar las características univariadas de simetría y dispersión, y las relaciones por pares entre variables.	92
4.9	Gráfico chi cuadrado; datos simulados presentados en el ejemplo con datos simulados del procedimiento 2 para evaluación de normalidad multivariada (página 94).	102
4.10	Gráfico chi cuadrado, datos pesos del corcho, tabla 4.2. . .	103
4.11	Gráfico Q-Q Beta para 500 observaciones $\frac{n}{(n-1)^2}T_i^2$ obtenidas por simulación de una distribución normal multivariada con $p = 3$	109

4.12	Histograma de los valores T_i^2 obtenidos de una muestra de 500 observaciones de una normal multivariada con $p = 3$. La curva superpuesta corresponde a la densidad estimada por suavizamiento kernel.	116
4.13	Carta de control T^2 en la Fase I, con 50 observaciones individuales simuladas. La observación produjo una señal; sin embargo, corresponde a un valor sin causa asignable, originado por la aleatoriedad.	119
4.14	Carta de control T^2 en la Fase II, con 20 observaciones individuales simuladas. El proceso aparece en control. . .	122
4.15	Carta de control T^2 en la Fase I, con 40 subgrupos simulados de tamaño 10. Los subgrupos 18, 24 y 30 aparecen arriba del UCL, pero no hay causa asignable.	124
4.16	Carta de control T^2 en la Fase II, con 20 nuevos subgrupos simulados de tamaño 10. El proceso aparece en control. . .	126
4.17	Gráficos de residuales para la regresión de X_3 vs. X_1, X_2, X_4, X_5 y X_6	132
4.18	Gráficos de residuales para la regresión de X_5 vs. X_1, X_2, X_3, X_4 y X_6	133
4.19	Serie simulada, su ACF y su PACF. La serie presenta una autocorrelación significativa, aunque es estacionaria y, según la ACF y la PACF, puede modelarse como un AR(1). . .	139
4.20	Gráfico de la serie z_t vs. sus rezagos, $k = 1, 2, \dots, 6$	140
4.21	Carta elipse al 5 % de significancia, para las dos primeras componentes, datos de la tabla 4.5. La observación 11 aparece fuera de control en la segunda componente.	155
4.22	Carta de control T^2 de las últimas tres componentes, datos de la tabla 4.5. La observación 13 aparece fuera de control. . .	158
4.23	Región de control elíptica y región de control definida por T_1^2 y $T_{2 1}^2$	165

4.24	Región de control elíptica y rectas de regresión relativas a las componentes condicionales $T_{1 2}^2$ y $T_{2 1}^2$	165
5.1	Probabilidad de señal para un <i>outlier</i> . Comparación de los métodos usual, DG y MVE. Donde ncp representa el parámetro de no centralidad.	186
5.2	Probabilidad de señal para dos <i>outliers</i> . Comparación de los métodos usual, DG y MVE.	187
5.3	Probabilidad de señal para tres <i>outliers</i> . Comparación de los métodos usual, DG y MVE.	187

Lista de tablas

2.1	Probabilidad de puntos fuera de límites de control.	17
2.2	Características eléctricas (dB) del ensamblaje final de once láminas de cerámica.	25
3.1	Números aleatorios de una distribución $N(\mu = 25, \sigma^2 = 9)$	48
4.1	Datos de absorción.	69
4.2	Pesos de corcho.	97
4.5	Cinco tipos de horas extras para el Departamento de Policía, Madison, Wisconsin.	152
4.6	Resumen descomposición MYT para ejemplo bivariado.	166



Prefacio

Este libro se concentra en la utilización del estadístico T^2 de Hotelling en el control estadístico de procesos multivariados, su implementación, sus propiedades y algunas de sus carencias, que son tema de investigación en la actualidad.

Se diseñó de manera que fuese autocontenido, en el sentido de que los elementos de control univariado necesarios para la comprensión de los principios básicos se incluyen de forma que la temática multivariada se pueda abordar sin ningún prerrequisito. Muchos de los procesos en la industria o en el sector de servicios dependen de diversas variables que usualmente están correlacionadas, y su correcto manejo aumenta la precisión en el control de procesos. El manejo univariado sigue siendo importante, pero es necesario tener en cuenta las características multivariadas, cuando sea del caso, y por ello se pretende dar en el texto una visión práctica que permita la implementación y el uso de las cartas de control T^2 multivariadas. Así las cosas, el texto va dirigido al usuario que quiera aprender a utilizar las técnicas, así como también a estudiantes avanzados de pregrado y de posgrado en Estadística e Ingeniería que pretendan hacer investigación aplicada en el área.

Se presenta la temática multivariada de manera amable, con instrucciones claras de implementación que contribuyan a la mejora de la calidad en nuestro medio. La experiencia docente de los autores en cursos de control de calidad con estadísticos e ingenieros, en las sedes de Bogotá y Medellín de la Universidad Nacional de Colombia, hacen de este trabajo un aporte importante a la consolidación de las maestrías profesionalizantes, en cuanto permite una relación más directa de la academia con el sector productivo. Es de anotar, también, que además del paquete SAS usado en los temas univariados, se presentan los códigos de los programas en R para las cartas de control T^2 . Este lenguaje es de código abierto y de gran utilidad, y por sus características y funcionalidad su uso es cada vez más universal en distintas áreas del saber. Se puede descargar sin costo desde <http://www.R-project.org>. Todos los códigos R usados en este libro se pueden obtener en el *link* <http://www.medellin.unal.edu.co/estadistica>, menú Nuestra Escuela - Cartas de Control T^2 Multivariadas Usando R y SAS.

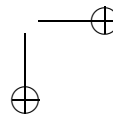
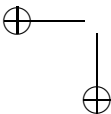
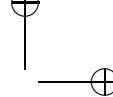
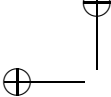
El libro se organizó en cinco capítulos. Los tres primeros presentan de manera rápida, pero precisa, los conceptos del control estadístico de procesos univariado, y sirven como introducción a los elementos fundamentales del área. El capítulo cuarto desarrolla en detalle las ideas básicas de la carta de control T^2 de Hotelling, la implementación de la carta de control y el chequeo en términos prácticos de los supuestos necesarios para su correcto uso; allí también se analiza el aspecto de detección de señales. Finalmente, el capítulo quinto muestra algunas variantes para robustecer la carta de control T^2 y mejorar su capacidad de detección en algunas circunstancias; se puede ver como una invitación a la investigación aplicada que se ubica en una de las fronteras de estado del arte del control multivariado de procesos.

Agradecemos a las Direcciones de Investigación de las sedes de Bogotá y Medellín de la Universidad Nacional de Colombia, a las Facultades de Ciencias y a los respectivos Departamentos o Escuelas de Estadística por la financiación y continua colaboración dentro de la filosofía de hacer de la Universidad un centro investigativo de excelencia. También queremos agradecer a los evaluadores y al Comité Editorial de la Facultad de Ciencias de la Sede de Bogotá por las valiosas sugerencias y correcciones que contribuyeron a mejorar el texto. Finalmente damos las gracias a las estadísticas Diana Pérez y María Carolina Paz por su colaboración en la edición de la primera y última versión de este libro, respectivamente.

Sergio Yáñez C.
Nelfi González Á.
Profesores, Escuela de Estadística, Universidad Nacional de Colombia,
sede Medellín

José A. Vargas N.
Profesor, Departamento de Estadística, Universidad Nacional de
Colombia, sede Bogotá

Medellín, Bogotá
Colombia
Octubre de 2011



CAPÍTULO 1

Introducción al control estadístico de procesos

1.1 Control estadístico de procesos (SPC)

El tema central de estas notas es el “control estadístico de procesos”, que se denotará SPC por la sigla en inglés universalmente reconocida y que abrevia la expresión *Statistical Process Control*. Con este nombre se reconoce la temática de control estadístico de calidad, que no es simplemente el uso de cartas de control, sino que abarca todo un sistema que permite desarrollar un mejoramiento continuo de la calidad.

1.2 Bosquejo histórico

Siguiendo a Vargas (2001), se presenta un rápido resumen histórico sobre control de calidad. La teoría estadística empieza a ser utilizada en control de calidad a partir de los años veinte. En 1924, Walter A. Shewhart, de Bell Telephone Laboratories, hizo el primer bosquejo de una carta de control, y en 1931 publicó el libro *Economic Control of*

6 CAPÍTULO 1. INTRODUCCIÓN AL CONTROL ESTADÍSTICO DE PROCESOS

Quality of Manufactured Product, en el cual se sentaron las bases del control estadístico de calidad. Harold F. Dodge y H. Roming, también de Bell System, publicaron en 1944 *Tablas de inspección por muestreo-muestreo simple y doble*. El trabajo de estas tres personas es la base de lo que hoy constituye el control estadístico de calidad.

A comienzos de los años cincuenta, W. Edwards Deming (1900-1993) desarrolló en Japón el concepto de calidad como un objetivo estratégico y económico, y mostró la forma de lograr tal objetivo. El impacto de Deming fue de tal magnitud, que la industria japonesa creció aceleradamente a partir de los años sesenta. Es así como en los setenta y comienzos de los ochenta se ve un resurgimiento de la utilización de las técnicas estadísticas con la participación activa de los países occidentales ante los buenos resultados japoneses. En un mercado abierto, los consumidores empiezan a exigir productos de buena calidad a un precio razonable. La industria toda, occidental y oriental, comprueba que la colocación de artículos de buena calidad en el mercado atrae compradores y, paralelamente, constituye a la larga un ahorro de dinero. La industria se embarca entonces en la implementación de procesos que la condujeron al mejoramiento de la calidad y la productividad, lo cual llevó a la filosofía de la calidad total (en sus diferentes variantes, como “calidad total de manejo” (TQM), “calidad total de compromiso” (TQI), “calidad total de excelencia” (TQE), entre otras) que requiere el tratamiento del desempeño en todos los aspectos de cualquier operación. Dentro de esta filosofía se enmarca el “control estadístico de procesos” (SPC).

Para terminar esta breve sinopsis histórica, es digno mencionar a Joseph M. Juran (1904), Eugene L. Grant (1897-1996) y George E. P. Box (1919) como pioneros insignes del trabajo en calidad (ver Ryan, 2000).

1.3 Algunas precisiones sobre el SPC

Como ya se mencionó, el acrónimo SPC representa a la expresión *Statistical Process Control*, que se tradujo como “control estadístico de procesos”. A la manera de Bissell (1994), se discutirá cada uno de los términos del acrónimo: control, proceso y estadística.

1.3.1 Control

La idea de control no es simplemente la de monitoreo; tiene aquí un sentido mucho más amplio de gestión del proceso o de gestión de calidad (*quality management*).

Los fundamentos de la mayoría de las filosofías de *gestión de calidad* se pueden resumir así:

Estrategia: planear y organizar para metas a largo plazo. El mejoramiento de la calidad debe ser dirigido por altos ejecutivos.

Logística: proporciona métodos y recursos para implementar la estrategia. Los recursos comprenderán entrenamiento y compromiso en tiempo de los altos ejecutivos para dirigir y apoyar el programa.

Desarrollo: presupone estar preparado para el cambio, para resolver problemas y para mejorar sistemas en todas las áreas de actividad.

Relaciones humanas: las personas son el recurso más valioso de la organización. Esto obliga a implementar una dirección participativa y la búsqueda de satisfacción en el trabajo.

SPC forma parte de la filosofía de la calidad total, la cual se debe extender a todos los aspectos del negocio. Se aplica a calidad y productividad de operaciones, a servicios tales como mantenimiento, transporte y suministro de materiales, y a áreas administrativas como seguridad, personal, gestión computacional y actividades de ventas.

1.3.2 El proceso

Normalmente, en SPC, la palabra ‘proceso’ está asociada con alguna forma de manufactura. Aquí, se tomará un punto de vista mucho más amplio, y se considerará un proceso como cualquier servicio manufacturero, administrativo, de papeleo en oficinas o cualquier otro sistema que corresponda a la siguiente secuencia:

Entrada → Actividad → Salida

8 CAPÍTULO 1. INTRODUCCIÓN AL CONTROL ESTADÍSTICO DE PROCESOS

Algunos ejemplos son:

Manuscrito → Procesamiento de palabras → Carta, documentos.

Investigación mercados → Actividad de eventos → Órdenes.

Órdenes → Selección de inventarios → Entrega.

Semillas, fertilizantes → Cultivo → Cosecha.

Materiales o componentes → Manufactura → Mercancía terminada.

Tradicionalmente el control de calidad ha tratado con detección de problemas, pero los autores insisten más en la prevención de problemas. Se trata así de buscar una estrategia que conduzca al resultado deseado. La figura 1.1 es un ejemplo sistémico de cómo lograr dicha estrategia.

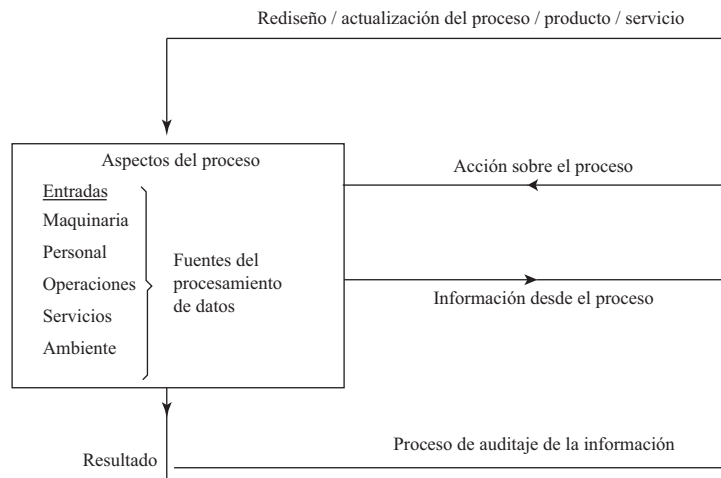


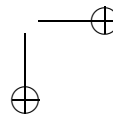
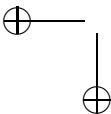
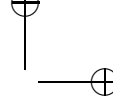
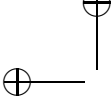
FIGURA 1.1 Búsqueda de una estrategia para un resultado deseado.

Ya sea en producción, diseño, administración o servicio, los resultados del control (más que detección de defectos) son operaciones más ágiles, reducción de costos, mayor producción, mejoramiento de la calidad, satisfacción del consumidor y mejores relaciones humanas. Este último resultado, identificado con la sensación de un trabajo bien hecho y con la satisfacción obtenida por las personas del conocimiento extraído de su experiencia junto con la posibilidad de actuar sobre él.

1.3.3 Estadística

En general, las técnicas estadísticas se necesitan para determinar si ha ocurrido variación anormal en lo que se está monitoreando, para detectar cambios en los parámetros del proceso y para identificar factores que están afectando las características del proceso. Estas notas tratan sobre algunas técnicas para lograr dichos objetivos, más allá de las temáticas de los cursos básicos de control. Incluirán, por ello, una discusión crítica de las tablas de control univariadas y se concentrarán en aplicaciones del análisis multivariado al control de calidad.

En síntesis, los métodos estadísticos deberán usarse (ver Ryan, 2000, p. 9) para identificar variaciones inusuales y señalar las causas de tales variaciones, ya sean del proceso de manufactura o debidos al negocio en general. El uso de los métodos estadísticos produce mejoras en la calidad, lo cual, a su vez, podría resultar en aumento de productividad.



CAPÍTULO 2

Cartas de control para mediciones con subgrupos

En este capítulo se analizarán las cartas de control que se pueden usar cuando se forman subgrupos (muestras) de datos, especialmente aquellas que suponen que las mediciones se pueden hacer con suficiente velocidad para permitir que los subgrupos se formen. Mediciones típicas son: longitud, ancho, diámetro, resistencia a la tensión y dureza de Rockwell.

2.1 Principios básicos de las cartas de control

Una carta (o gráfico) de control es un gráfico de los datos contra el tiempo u orden de producción, y se usa principalmente para el estudio y control de procesos repetitivos. La figura 2.1 es un ejemplo de una carta de control típica donde las siglas de los límites de control se escriben en inglés, pues es la forma como aparecen en el paquete SAS. Usualmente se les denomina cartas de control Shewhart en honor a su creador.

Se pueden controlar variables del proceso (por ejemplo, temperatura y presión) o variables de producto (por ejemplo, diámetros y espesores).

12 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

En general es deseable monitorear todas las variables del proceso que afecten importantes variables del producto (ver Ryan, 2000, pp. 77-78).

Las cartas de control no pueden producir, por sí solas, control estadístico; este es el trabajo del personal responsable del proceso. Las cartas de control pueden indicar si el control estadístico se está manteniendo y suministra a los usuarios señales adicionales de los datos.

La base de la teoría de las cartas de control depende de la variabilidad en la calidad de un producto. Se pueden distinguir dos tipos de variabilidad: la primera debida a “causas aleatorias”, es decir, a las variaciones naturales del proceso, sobre las cuales poco se puede hacer, y la segunda debida a las “causas asignables” (por ejemplo, una máquina descalibrada) sobre las cuales se puede actuar.

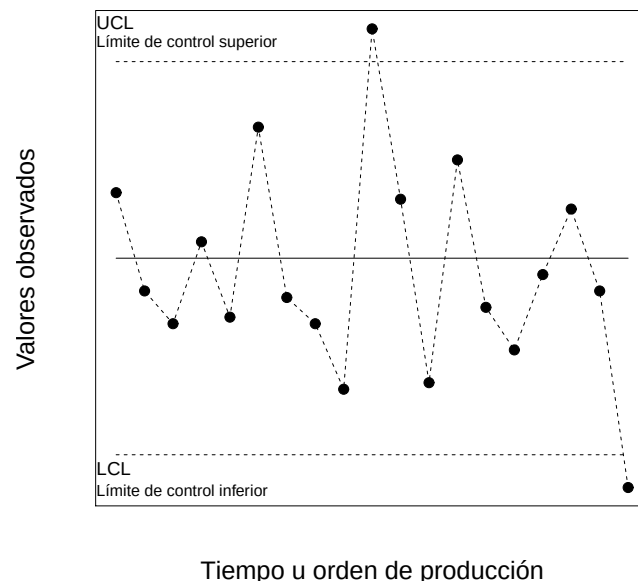


FIGURA 2.1 Carta de control típica.

Por otro lado, si los datos se mantienen dentro de un rango preestablecido, se dice que el proceso está *bajo control*. Si, por otra parte, se presenta variación debida a una o varias causas asignables, se dice que el proceso está *fuera de control*. La figura 2.2 presenta un esquema general del objetivo de una carta de control.

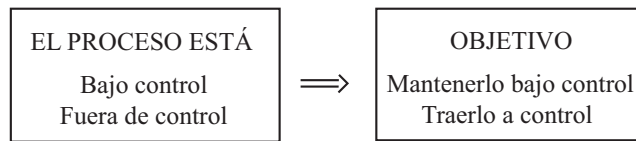


FIGURA 2.2 Objetivo de una carta de control.

El uso de una carta de control requiere generalmente dos fases (ver Vargas, 2001). En la Fase I se analiza un conjunto histórico de datos con el objetivo de saber si el proceso estaba bajo control. Para hacerlo se recomienda obtener al menos 20 subgrupos o mínimo 100 observaciones individuales (dependiendo de si se van a usar subgrupos u observaciones individuales). Con base en estos datos se hace una primera estimación de los parámetros desconocidos del proceso y se establecen unos límites de control iniciales (como el UCL y el LCL de la figura 2.1). Se grafican los datos observados que corresponden a los puntos en la carta, y si uno o más puntos quedan ubicados por fuera de los límites se buscan las causas asignables, y si se pueden remover, se eliminan los puntos y se recalculan los límites de control. Si la causa no puede ser removida, se debe ver como parte permanente del proceso, de manera que los límites iniciales no deberían recalcularse. Este proceso de recálculo continúa hasta que se tengan puntos fuera de límites a los que se les pueda detectar causa asignable y dicha causa pueda ser removida. Después de que el proceso se considere en control se pasa a la Fase II, una fase de monitoreo con el objetivo de mantener el proceso bajo control.

2.2 Tipos de errores (riesgos del muestreo)

Considérese el caso en que se tomen muestras pequeñas de un proceso a intervalos regulares de tiempo, y se construya una carta de control para la media y el rango (estas cartas se tratarán en detalle en la sección 2.4 de este capítulo); como resultado, se concluye que el proceso está bajo control o fuera de control. Si el proceso está fuera de control, podría ser debido a un cambio en el nivel promedio o en la variabilidad del proceso. En razón a la variación inherente del muestreo, los promedios y rangos varían de muestra a muestra, aunque la media y el rango reales del proceso sean constantes; esto da lugar a dos tipos de riesgos.

14 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

Riesgos de tipo I: El riesgo de que una muestra caiga fuera de límites cuando no ha ocurrido cambio en el proceso.

Riesgos de tipo II: El riesgo de que una muestra caiga dentro de los límites de control, aunque un cambio real haya ocurrido en el proceso.

Es usual darle prioridad a un riesgo de tipo I. A este riesgo se le puede denominar como tasa de “falsas alarmas”. Una falsa alarma se produce cuando un punto cae por fuera de los límites de control debido solamente a causas aleatorias. Así, por ejemplo, una tasa de falsas alarmas del 5 % ($\alpha = 0,05$) quiere decir que uno de cada 20 subgrupos produce una falsa alarma ($\frac{1}{20} = 0,05$). En control de calidad es usual tomar un $\alpha = 0,005$, lo cual diría que uno de cada 200 subgrupos produce una falsa alarma ($\frac{1}{200} = 0,005$).

En general, el tamaño de muestra está determinado por la tasa de producción, y variará de periodo a periodo. El tamaño de la muestra y la frecuencia del muestreo deben ser determinados de manera conjunta con los directivos de la empresa. Lo ideal sería tomar muestras grandes a intervalos cortos para mejor protección contra desvíos en el proceso, pero resulta costoso. Así, el problema práctico es, entonces, cuándo tomar muestras grandes a intervalos menos frecuentes o muestras pequeñas a intervalos más frecuentes.

Estos tipos de errores son similares a los que se definen en una prueba de hipótesis. Por esta razón, algunos autores expresan que los procedimientos de cartas de control y pruebas de hipótesis son equivalentes o por lo menos están muy relacionados. Sin embargo, se debe tener cuidado con esta equivalencia. La aplicación de las cartas de control en Fase II, en la cual se asume conocida la distribución de la característica de calidad y los parámetros bajo control, es muy parecida al proceso repetitivo de prueba de hipótesis. Pero en Fase I no hay equivalencia de los dos métodos, pues los objetivos y el manejo de las cartas de control en Fase I son diferentes a los de la Fase II. Woodall (2000) presenta una discusión interesante acerca de este tema.

2.3 Aspectos estadísticos básicos de las cartas de control

Si el interés está centrado en controlar la media del proceso, μ , y los límites se dan como $\mu \pm 3\sigma_{\bar{x}}$, bajo el supuesto de que X tiene una distribución normal con $\sigma_{\bar{x}}$ conocida, se tiene una probabilidad total fuera de límites de

$$1 - P[\mu - 3\sigma_{\bar{x}} \leq \bar{x} \leq \mu + 3\sigma_{\bar{x}}] = 1 - 0,9973 = 0,0027 \text{ (esto es 0,00135 a cada lado)}$$

lo que significa que habrá 27 posibilidades de 10000 de observar un valor fuera de límites para $\bar{x}$ cuando la media está en μ (ver figura 2.3).

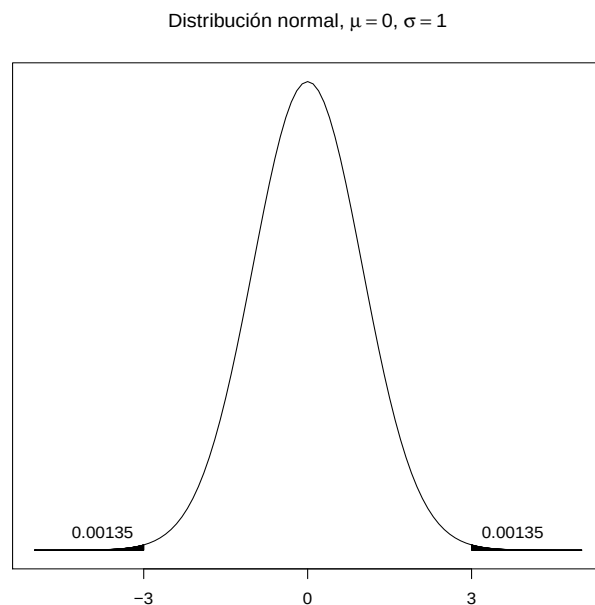


FIGURA 2.3 Distribución normal.

Ahora bien, en la práctica no se tienen, generalmente, distribuciones normales, ni tampoco se conocen la media verdadera del proceso μ , ni el valor verdadero de $\sigma_{\bar{x}}$. Por tanto, los “límites 3-sigma” (como usualmente se les conoce) no son límites probabilísticos puesto que las probabilidades exactas son desconocidas. Esto es importante para entender lo que

16 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

los ampliamente usados “límites 3-sigma” realmente significan. Incluso si se conociera μ , no se puede esperar que se mantenga constante en un periodo largo de tiempo. De esta manera, cuando estas probabilidades se aplican al futuro, son solo aproximaciones.

No obstante, si se tienen subgrupos (muestras) de tamaño al menos 4 o 5, la distribución de $\bar{x}$ no diferirá mucho de la distribución normal, siempre y cuando la distribución de X no se aleje mucho de la distribución normal. Esto resulta del hecho de que la distribución de $\bar{x}$ será más normal que la distribución de X (Ryan, 2000). Aun si la distribución es bastante asimétrica, se pueden, usualmente, transformar los datos (por ejemplo, log, raíz cuadrada, recíproca) de forma que se obtengan datos aproximadamente normales.

2.3.1 Cartas de control en tiempo real y sobre datos históricos

Es pertinente detenerse a analizar la determinación de los límites de control en la Fase I. Cuando un conjunto de puntos se grafican todos al mismo tiempo (en la Fase I y tal vez incluso en la Fase II), la probabilidad de observar al menos un punto fuera de los límites de control será, obviamente, mucho mayor que 0,0027, el cual aplica a puntos graficados individualmente con límites 3-sigma distribución normal y parámetros conocidos.

Para n puntos, la probabilidad de tener al menos una observación fuera de límites de control 3-sigma se calcularía como:

Sean

p = probabilidad de tener una observación fuera de límites de control con una sola observación.

$q = 1 - p$ = probabilidad de que ninguna observación caiga fuera de límites de control.

Se sabe que $p = 0,0027$ y $q = 0,9973$ bajo los supuestos arriba mencionados. Así, considerando un esquema binomial con n como el número de ensayos, se tiene

$$P(\geq 1 \text{ puntos fuera de límites} \mid n \text{ puntos}) = 1 - q^n \quad (2.1)$$

2.3. ASPECTOS ESTADÍSTICOS BÁSICOS DE LAS CARTAS DE CONTROL 17

En la tabla 2.1 se tienen las probabilidades verdaderas, calculadas con la ecuación (2.1), para distintos valores de n y una aproximación dada por $(0,0027)n$ que se comporta bastante bien para valores moderados de n .

TABLA 2.1 Probabilidad de puntos fuera de límites de control.

n	0,0027n	Probabilidad real (≥ 1 puntos fuera de límites)
1	0,0027	0,0027
2	0,0054	0,0054
5	0,0135	0,0134
10	0,0270	0,0267
15	0,0405	0,0397
20	0,0540	0,0526
25	0,0675	0,0654
50	0,1350	0,1264
100	0,2700	0,2369
350	0,9450	0,6118

Fuente: Ryan (2000), p. 79.

Lo importante de resaltar es que la probabilidad de observar al menos un punto fuera de límites cuando, digamos, 15 o 20 puntos se grafican simultáneamente, es mucho mayor que la probabilidad cuando se grafica un solo punto. Cuando los puntos se grafican individualmente en tiempo real, la probabilidad de 0,0027 aplica a cada punto de manera que si el proceso está en control hay, en efecto, una probabilidad muy pequeña de que un punto particular caiga fuera de límites. Pero cuando se determinan los límites de control tentativos y se revisan periódicamente (usando un conjunto de observaciones cada vez) y cuando el monitoreo con cartas de control no se realiza en tiempo real, la probabilidad de observar uno o más puntos fuera de límites, dado el proceso en control, es obviamente mucho mayor.

Esto no quiere decir que se deberían ignorar tales puntos y no indagar por las causas asignables; simplemente, que no es motivo de asombro no encontrar dichas causas asignables.

Aunque el uso de límites 3-sigma se ha vuelto lo usual, al menos en las aplicaciones manufactureras, no hay razón para que se usen siempre. Si, por ejemplo, una situación particular exige que 20 puntos sean graficados

18 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

simultáneamente en una carta cada vez que se hace el control, los límites se pueden ajustar de forma que, si se quiere, la probabilidad de observar al menos 1 de los 20 puntos fuera de límites sea cercana a 0,0027 cuando el proceso está en control.

De la regla práctica de la tabla 2.1 se usaría

$$np = 0,0027$$

de manera que con $n = 20$

$$20p = 0,0027$$

$$p = 0,000135$$

Buscando en la tabla de la distribución normal, corresponde aproximadamente a $Z = 3,81$ (repartiendo 0,0000675 en cada cola). Entonces se podrían límites usar 3,81-sigma. Esto no es para sugerir que *debería* hacerse, sino que *podría* hacerse.

Como se anotó antes, este análisis aplica solamente para el caso en que se supone que los parámetros se conocen. Cuando los parámetros son desconocidos, la probabilidad verdadera de que al menos uno de n puntos sean graficados fuera de límites cuando el proceso está en control no se puede determinar analíticamente puesto que las desviaciones de los n puntos de los límites de control están correlacionadas, ya que cada desviación contiene realizaciones de variables aleatorias comunes. Esto se cumple tanto para la Fase I como para la Fase II (ver Sullivan y Woodall, 1996).

En consecuencia, una probabilidad exacta no resultaría de cálculos tales como los del último ejemplo. Las probabilidades exactas solo se pueden determinar por simulación.

2.4 Construcción de cartas de control

Las cartas $\bar{X}$ han sido las cartas de control más usadas. Pero antes de utilizarlas se recomienda tener en estado de control estadístico la variabilidad del proceso, pues de otra manera no se podría tener una distribución estable de mediciones con una sola media fija. Por ello se

describirán primero las cartas de control R , s y s^2 diseñadas para controlar la variabilidad del proceso.

Se sabe que se puede estandarizar $\bar{X}$ así

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

con

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

de tal manera que $Z \sim N(0, 1)$ si $X \sim N(\mu, \sigma^2)$. La media μ del proceso, generalmente desconocida, se estima por $\bar{\bar{x}}$ donde

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \bar{x}_i}{k}$$

para los promedios de k subgrupos. La desviación estándar del proceso σ se puede estimar usando s (la desviación estándar del proceso) o R (el rango). Los promedios de s o de los rangos R no son estimadores insesgados de σ ; para ello se han construido tablas de constantes de tal forma que al dividir los citados promedios por dichas constantes se obtengan estimadores insesgados. Dichas constantes y otras que se usarán en las distintas cartas de control se encuentran tabuladas en varios libros de control de calidad, por ejemplo Ryan (2000, p. 540, tabla E), para diferentes tamaños de subgrupos. Así, por ejemplo, si se usan rangos, se tendrá

$$\hat{\sigma} = \frac{\bar{R}}{d_2}$$

donde $\bar{R}$ es el promedio de los rangos de cada subgrupo.

Si se usan desviaciones estándar, se tendrá

$$\hat{\sigma} = \frac{\bar{s}}{c_4}$$

donde $\bar{s}$ es el promedio de las desviaciones estándar de los subgrupos. En la citada tabla se encuentran los valores de d_2 y c_4 . Las derivaciones teóricas relativas a estas constantes se pueden encontrar en Ryan (2000, pp. 123-125), al cual referimos al lector interesado.

2.4.1 Carta R

Considérense límites 3-sigma para la carta R, así como para todas las cartas de control “estándar” que se presentarán. Así, se tienen los límites de control para R como

$$\bar{R} \pm 3\hat{\sigma}_R \quad (2.2)$$

los límites dados por la ecuación (2.2), se puede mostrar, son iguales a:

$$\begin{aligned} LCL &= \bar{R} - 3\hat{\sigma}_R = D_3\bar{R} \\ UCL &= \bar{R} + 3\hat{\sigma}_R = D_4\bar{R} \end{aligned}$$

Valores de D_3 y D_4 para distintos tamaños muestrales, cabe recordar, se pueden encontrar en la mencionada tabla E de Ryan (2000). Estos valores dependen del supuesto de normalidad de las observaciones individuales. En la práctica, la distribución de R es bastante asimétrica, luego no es cercana a la distribución normal; sin embargo, cuando se usan límites 3-sigma se aceptan los supuestos.

La aproximación estadística formal sería no utilizar los límites 3-sigma, sino “límites probabilísticos”, que calculan de manera más exacta los límites de control teniendo en cuenta la forma distribucional del rango, que es altamente asimétrica. Para ello hay tablas, como la que se puede encontrar en Ryan (2000, p. 541, tabla F) y en Harter (1960).

2.4.2 Carta R con límites probabilísticos

Los límites de control probabilísticos para la carta R se obtienen así:

$$\begin{aligned} LCL &= D_{\alpha/2} \hat{\sigma}_R \\ UCL &= D_{1-\alpha/2} \hat{\sigma}_R \end{aligned}$$

donde $\hat{\sigma}_R = \bar{R}/d_2$ se reparte en colas iguales (suponiendo normalidad). Así, por ejemplo, $D_{0,001}$ y $D_{0,999}$ corresponden a $\alpha = 0,002$, donde cada cola tiene un área de 0,001. En este caso, de la tabla F mencionada arriba se tiene para $n = 4$ subgrupos $D_{0,001} = 0,199$ y $D_{0,999} = 5,309$.

¿Cuál conjunto de límites se debería usar en la práctica? Ambos conjuntos se basan en el supuesto de normalidad (de X) y en el supuesto de

$\sigma = \bar{R}/d_2$. Ninguno de los supuestos se cumple usualmente en la práctica, y además la teoría estadística no existe para permitir una comparación de las dos metodologías bajo diferentes condiciones (en particular, la distribución del rango no es muy conocida y en especial no está tabulada para distribuciones distintas de la normal). Los límites probabilísticos son atractivos desde el punto de vista estadístico, en la medida en que intentan corregir la asimetría de la distribución del rango. Sin embargo, dado el estado del arte de la temática, los más usados son todavía los límites 3-sigma.

2.4.3 Carta s

Los límites de control para la carta de control s se obtienen así:

$$\bar{s} \pm 3\hat{\sigma}_s \quad (2.3)$$

donde $\bar{s}$ es el promedio de las desviaciones estándar de los subgrupos y $\hat{\sigma}_s$ es el estimativo de la desviación estándar de s .

Los límites dados por la ecuación (2.3), se puede mostrar, son iguales a:

$$\begin{aligned} LCL &= \bar{s} - 3\hat{\sigma}_s = B_3\bar{s} \\ UCL &= \bar{s} + 3\hat{\sigma}_s = B_4\bar{s} \end{aligned}$$

donde B_3 y B_4 se pueden encontrar en la tabla ya mencionada. Como en el caso de la carta R , estos límites se basan en el supuesto de normalidad.

2.4.4 Carta s con límites probabilísticos

Si $X \sim N(\mu, \sigma^2)$, entonces

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

De aquí se tiene

$$P\left(\chi_{\alpha/2;n-1}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{1-\alpha/2;n-1}^2\right) = 1 - \alpha$$

22 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

de donde

$$P \left(\sigma \sqrt{\frac{\chi_{\alpha/2;n-1}^2}{n-1}} < s < \sigma \sqrt{\frac{\chi_{1-\alpha/2;n-1}^2}{n-1}} \right) = 1 - \alpha$$

Así, si la variabilidad del proceso está en control en σ , $(1 - \alpha)$, 100 % de las veces la desviación estándar de los subgrupos caerá entre los extremos del intervalo.

De esta manera se obtienen los siguientes límites de control probabilísticos:

$$LCL = \hat{\sigma} \sqrt{\frac{\chi_{\alpha/2;n-1}^2}{n-1}}$$

$$UCL = \hat{\sigma} \sqrt{\frac{\chi_{1-\alpha/2;n-1}^2}{n-1}}$$

donde $\hat{\sigma} = \bar{s}/c_4$ y la línea central de la carta estaría en $\bar{S}$.

Se pueden hacer observaciones similares a las de las cartas R , concretamente sobre los supuestos y sobre las ventajas estadísticas de los límites probabilísticos. En este punto es pertinente anotar que las cartas de control o cualquier otro procedimiento estadístico no se invalidan porque los supuestos no se cumplan exactamente en la práctica. Lo que importa determinar es qué tan insensibles son a la violación de los supuestos. Algunos estudios (ver Ryan, 2000, p. 114) muestran la robustez de las cartas R y $\bar{X}$ a la no normalidad, a menos que haya una gran desviación de la normalidad. Las cartas s^2 y s son más sensibles a desviaciones leves y moderadas de la normalidad. En estos últimos casos se puede intentar transformar los datos de manera que se aproximen a una distribución normal y luego aplicar los procedimientos básicos de control a la variable transformada.

2.4.5 Carta s^2

Con las cartas s^2 se puede controlar la varianza del proceso. En este caso los límites de control probabilísticos son:

$$LCL = \bar{s}^2 \left(\frac{\chi_{\alpha/2;n-1}^2}{n-1} \right)$$

$$UCL = \bar{s}^2 \left(\frac{\chi_{1-\alpha/2; n-1}^2}{n-1} \right)$$

Obsérvese que los límites son diferentes de los de las cartas s , puesto que $\bar{s}^2$ es un estimador insesgado de σ^2 mientras que $\left(\bar{s}/c_4\right)^2$ no es insesgado, donde $\bar{s}^2$ es el promedio de los s^2 de cada subgrupo.

2.4.6 Carta $\bar{X}$

Una vez la variabilidad del proceso se pueda considerar en estado de control estadístico, se puede proceder a investigar si la media está o no en control. Para tal propósito se usará una carta $\bar{X}$. Los límites de control para una carta $\bar{X}$ se obtienen así:

$$\bar{\bar{x}} \pm 3\hat{\sigma}_{\bar{x}}$$

donde $\bar{\bar{x}}$ denota el promedio global de los promedios de los subgrupos y $\sigma_{\bar{x}}$ denota un estimador de la desviación estándar de los promedios de los subgrupos.

Se sabe que

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

por tanto

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$$

Según el estimador escogido para $\hat{\sigma}_x$, se tendrán distintas cartas, así

1. Carta $\bar{X} - R$

Si se emplea R para estimar σ

$$\hat{\sigma} = \bar{R}/d_2$$

los límites de control son:

$$\bar{\bar{x}} \pm 3\hat{\sigma}_{\bar{x}} = \bar{\bar{x}} + \frac{3\hat{\sigma}_x}{\sqrt{n}} = \bar{\bar{x}} \pm \frac{3(\bar{R}/d_2)}{\sqrt{n}} = \bar{\bar{x}} \pm A_2\bar{R}$$

24 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

donde

$$A_2 = \frac{3}{(d_2\sqrt{n})}$$

Estos valores de A_2 se pueden encontrar en la tabla antes mencionada.

2. Carta $\bar{X} - s$

Si usamos s en lugar de R para estimar σ , los límites de control son:

$$\bar{\bar{x}} \pm \frac{3 \left(\bar{s}/c_4 \right)}{\sqrt{n}} = \bar{\bar{x}} \pm A_3 \bar{s}$$

donde

$$A_3 = \frac{3}{(c_4\sqrt{n})}$$

Estos valores de A_3 se pueden encontrar en la tabla antes mencionada.

En la práctica, si se usa una carta R o una carta s , se deberían usar cartas $\bar{X} - R$ o cartas $\bar{X} - s$ para controlar la media.

Se sabe también que para tamaños de subgrupos de hasta 5, los estimadores R y s se comportan de manera similar. Pero para muestras grandes, tamaño de subgrupo ≥ 10 , se prefiere a s , pues el rango es mucho más sensible a la ocurrencia de un valor extremo (ver SAS/QC, 1999, p. 1383).

2.5 Análisis en SAS usando PROC SHEWHART

Se ilustrará el uso de las cartas de control para subgrupos con los datos de la tabla 2.2.

Los datos se recopilaron con el objetivo de determinar si la variabilidad de una característica eléctrica particular en el ensamblaje de ciertas unidades electrónicas era significativa, al comparar once láminas cerámicas relativas a la variabilidad entre siete franjas dentro de cada lámina. Si la variabilidad entre las láminas resultara significativa, se descartarían las láminas inferiores.

TABLA 2.2 Características eléctricas (dB) del ensamblaje final de once láminas de cerámica.

16,5	15,7	17,3	16,9	15,5	13,5	16,5	16,5	14,5	16,9	16,5
17,2	17,6	15,8	15,8	16,6	13,5	14,3	16,9	14,9	16,5	16,7
16,6	16,3	16,8	16,9	15,9	16,0	16,9	16,8	15,6	17,1	16,3
15,0	14,6	17,2	16,8	16,5	15,9	14,6	16,1	16,8	15,8	14,0
14,4	14,9	16,2	16,6	16,1	13,7	17,5	16,9	12,9	15,7	14,9
16,5	15,2	16,9	16,0	16,2	15,2	15,5	15,0	16,6	13,0	15,6
15,5	16,1	14,9	16,6	15,7	15,9	16,1	16,1	10,9	15,0	16,8
$\bar{x} = 16,0$	15,8	16,4	16,5	16,1	14,8	15,9	16,3	14,6	15,7	15,8
$R = 2,8$	3,0	2,4	1,1	1,1	2,5	3,2	1,9	5,9	4,1	2,8

Fuente: Ryan (2000, p. 99).

Para el análisis se usará el PROC SHEWHART de SAS. Se comienza con las cartas $\bar{X} - R$ (en las gráficas se denotan como Carta Xbar-R). Abajo aparece el programa con sus resultados. La gráfica superior es la carta $\bar{X}$ y la inferior es la R. Las tablas numéricas corresponden a los valores de cada carta.

Código SAS 2.1

```
options ls=80 ps=60 nodate nocenter nonumber;
options colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
TITLE 'Carta Xbar-R';
symbol v=dot;
PROC Shewhart data=sasuser.tabla4_4;
xrchart y*subg/tableall; run;
```

Salida SAS 2.1

Carta Xbar-R
Procedimiento SHEWHART

Resumen del diagrama de medias y rangos para y

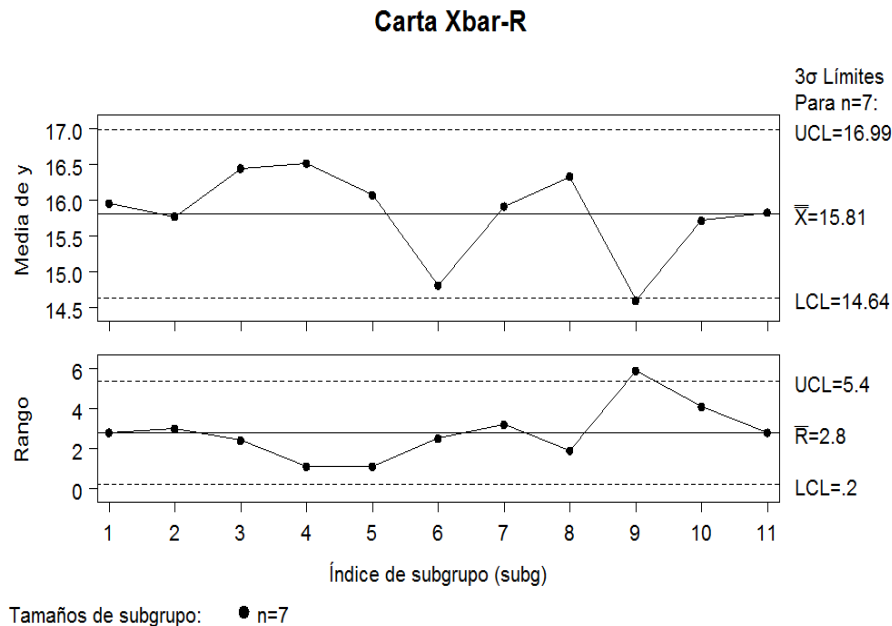
		-----Límites 3 Sigma con n=7 para la media-----				
subg	Tamaño de la muestra del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.640291	15.957143	15.814286	16.988281	
2	7	14.640291	15.771429	15.814286	16.988281	
3	7	14.640291	16.442857	15.814286	16.988281	
4	7	14.640291	16.514286	15.814286	16.988281	
5	7	14.640291	16.071429	15.814286	16.988281	
6	7	14.640291	14.814286	15.814286	16.988281	
7	7	14.640291	15.914286	15.814286	16.988281	
8	7	14.640291	16.328571	15.814286	16.988281	
9	7	14.640291	14.600000	15.814286	16.988281	Inferior
10	7	14.640291	15.714286	15.814286	16.988281	
11	7	14.640291	15.828571	15.814286	16.988281	

Resumen del diagrama de medias y rangos para y

		-----Límites 3 Sigma con n=7 para el rango-----			
subg	Límite inferior	Rango subgrupo	Rango Media	Límite superior	Límite sobrepasado

26 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

1	0.21198168	2.80000000	2.80000000	5.3880183	
2	0.21198168	3.00000000	2.80000000	5.3880183	
3	0.21198168	2.40000000	2.80000000	5.3880183	
4	0.21198168	1.10000000	2.80000000	5.3880183	
5	0.21198168	1.10000000	2.80000000	5.3880183	
6	0.21198168	2.50000000	2.80000000	5.3880183	
7	0.21198168	3.20000000	2.80000000	5.3880183	
8	0.21198168	1.90000000	2.80000000	5.3880183	
9	0.21198168	5.90000000	2.80000000	5.3880183	Superior
10	0.21198168	4.10000000	2.80000000	5.3880183	
11	0.21198168	2.80000000	2.80000000	5.3880183	



De la carta R se observa que el subgrupo 9 excede el UCL, lo cual también se puede notar de la tabla numérica del rango en la línea del subgrupo 9, donde aparece el mensaje de límite excedido. También en la carta $\bar{X}$ se puede observar que la lámina 9 está por debajo del LCL, lo cual también se advierte en la correspondiente tabla numérica. Así, se ve que hay aparentemente un problema con el subgrupo 9 en cuanto a variabilidad y promedio. Un análisis detallado revela que para esta lámina hay un valor de 10,9, mucho menor que los otros valores de la tabla. Los encargados de calidad en la empresa consideran que dicho dato es quizás un *outlier* o un dato registrado incorrectamente; por ello se asume que hay causa asignable y se procede a eliminarlo y a recalcular los límites para ambas cartas de control. Se recomienda, en la práctica, hacer estos recálculos conjuntos (ver Ryan, 2000). A continuación aparece el programa SAS para hacer este cálculo con sus resultados.

Código SAS 2.2

```
options ls=80 ps=60 nodate nocenter nonumber;
goptions colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
Data Calidad.tabla4_4M;
set calidad.tabla4_4;
where y<=>10.9; run;
TITLE 'Carta Xbar-R';
symbol v=dot;
PROC Shewhart data=calidad.tabla4_4M;
xrchart y*subg/tableall; run;
```

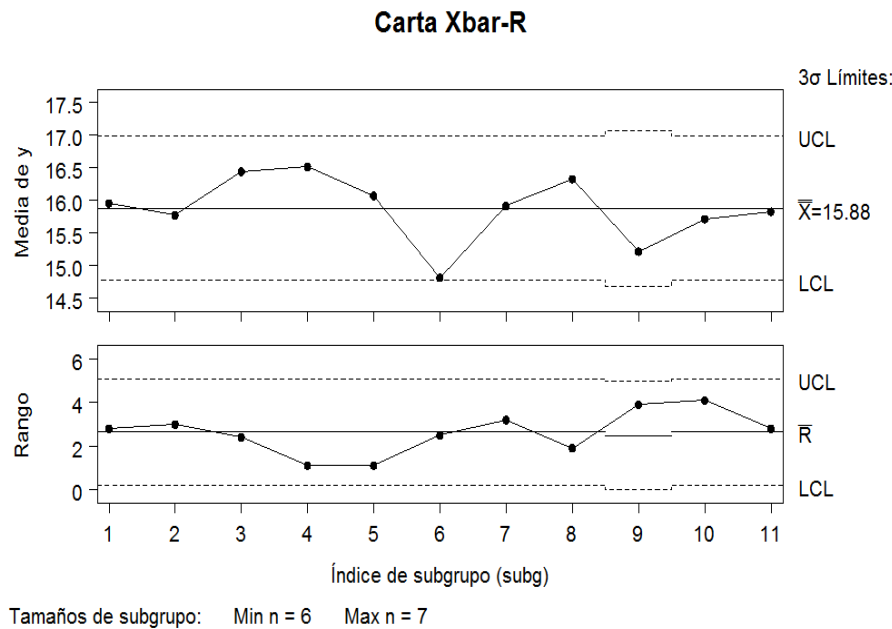
Salida SAS 2.2

Carta Xbar-R
Procedimiento SHEWHART

Resumen del diagrama de medias y rangos para y						
-----Límites 3 Sigma para la media-----						
subg	Tamaño de la muestra del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.771218	15.957143	15.878947	16.986677	
2	7	14.771218	15.771429	15.878947	16.986677	
3	7	14.771218	16.442857	15.878947	16.986677	
4	7	14.771218	16.514286	15.878947	16.986677	
5	7	14.771218	16.071429	15.878947	16.986677	
6	7	14.771218	14.814286	15.878947	16.986677	
7	7	14.771218	15.914286	15.878947	16.986677	
8	7	14.771218	16.328571	15.878947	16.986677	
9	6	14.682463	15.216667	15.878947	17.075432	
10	7	14.771218	15.714286	15.878947	16.986677	
11	7	14.771218	15.828571	15.878947	16.986677	

Resumen del diagrama de medias y rangos para y					
-----Límites 3 Sigma para el rango-----					
subg	Límite inferior	Rango subgrupo	Media	Límite superior	Límite sobrepasado
1	0.20001650	2.8000000	2.6419557	5.0838949	
2	0.20001650	3.0000000	2.6419557	5.0838949	
3	0.20001650	2.4000000	2.6419557	5.0838949	
4	0.20001650	1.1000000	2.6419557	5.0838949	
5	0.20001650	1.1000000	2.6419557	5.0838949	
6	0.20001650	2.5000000	2.6419557	5.0838949	
7	0.20001650	3.2000000	2.6419557	5.0838949	
8	0.20001650	1.9000000	2.6419557	5.0838949	
9	0.00000000	3.9000000	2.4759330	4.9613484	
10	0.20001650	4.1000000	2.6419557	5.0838949	
11	0.20001650	2.8000000	2.6419557	5.0838949	

Se podría haber excluido el subgrupo 9 entero, pero por el análisis anterior no fue necesario. Además, en SAS se tiene la opción de trabajar con subgrupos de tamaño variable, la cual se irá explorando durante el análisis del ejemplo.



Observando las cartas y las tablas numéricas, se observa que los resultados aparecen en control. La carta R muestra una variabilidad controlada. En la carta $\bar{X}$, si bien tiene todos los puntos dentro del límite, se observa que el subgrupo 6 está muy cerca del límite inferior, por lo cual se podría considerar que esta lámina es inferior a las otras; más adelante se hará este recálculo.

Por ahora obsérvese en las cartas y las tablas numéricas que los límites de control para el subgrupo 9 son diferentes debido a que el tamaño de este subgrupo es ahora de 6 al excluir el dato 10,9. Para ello SAS utiliza los siguientes límites de control que tienen en cuenta el tamaño del subgrupo:

Carta $\bar{X}$:

$$LCL = \bar{\bar{X}} - 3\hat{\sigma} / \sqrt{n_i}$$

$$UCL = \bar{\bar{X}} + 3\hat{\sigma} / \sqrt{n_i}$$

donde n_i es el tamaño del i -ésimo subgrupo.

Carta R:

$$LCL = \max(d_2(n_i)\hat{\sigma} - 3d_3(n_i)\hat{\sigma}, 0)$$

$$UCL = d_2(n_i)\hat{\sigma} + 3d_3(n_i)\hat{\sigma}$$

donde

$d_2(n)$: el valor esperado del rango de n variables normales independientemente distribuidas con desviación estándar uno.

$d_3(n)$: el error estándar del rango de n observaciones independientes de una población normal con desviación estándar uno.

Obsérvese también que la línea control de la carta R cambia con el tamaño del subgrupo, pues se calcula como $d_2(n_i)\hat{\sigma}$.

Ahora bien, para calcular $\hat{\sigma}$ se utilizan las siguientes metodologías:

Método usual

Es el presentado en secciones anteriores, pero teniendo en cuenta el tamaño del subgrupo, así:

$$\hat{\sigma} = \frac{R_1/d_2(n_1) + \cdots + R_N/d_2(n_N)}{N}$$

donde N es el número de subgrupos y R_i es el rango del i -ésimo subgrupo.

Método MVLUE

MVLUE se refiere a la sigla en inglés *Minimum Variance Linear Unbiased Estimate*, que es estimador de mínima varianza insesgado y se calcula así:

$$\hat{\sigma} = \frac{f_1 R_1/d_2(n_1) + \cdots + f_N R_N/d_2(n_N)}{f_1 + \cdots + f_N}$$

donde

$$f_i = \frac{[d_2(n_i)]^2}{[d_3(n_i)]^2}$$

30 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

El MVLUE asigna mayor peso a estimativos de σ de subgrupos con mayor tamaño muestral; por ello se recomienda en situaciones en las cuales los tamaños muestrales de los subgrupos varían. Si los tamaños de los subgrupos son constantes, el MVLUE se reduce al método usual.

A manera de ilustración se calculará el MVLUE para el ejemplo citado. En el programa SAS se puede ver la opción SMETHOD = MVLUE.

Código SAS 2.3

```
options ls=80 ps=60 nodate nocenter nonumber;
goptions colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
Data Calidad.tabla4_4M;
set calidad.tabla4_4;
where y<>10.9;
run;
TITLE 'Carta Xbar-R';
symbol v=dot;
PROC Shewhart data=calidad.tabla4_4M;
xrchart y*subg/smethod=mvlu tableall;
run;
```

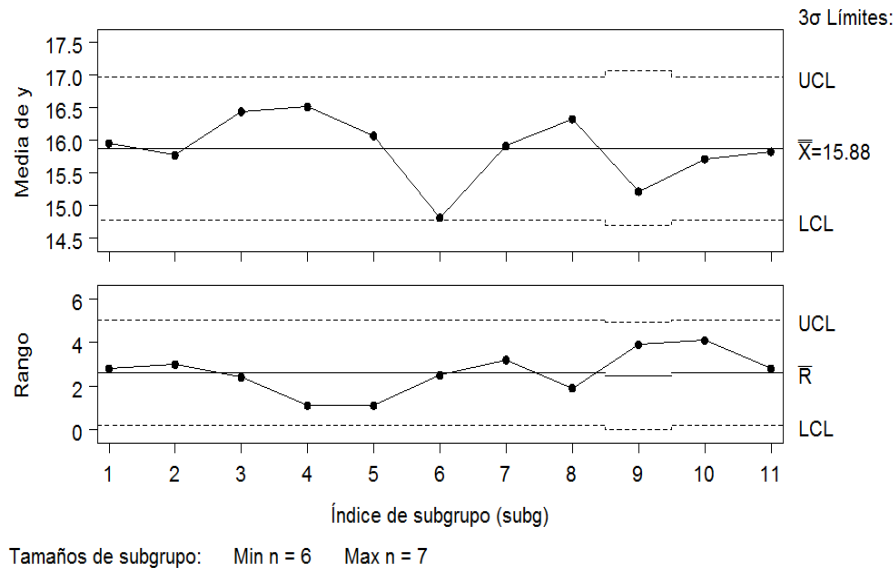
Salida SAS 2.3

Carta Xbar-R
Procedimiento SHEWHART

Resumen del diagrama de medias y rangos para y						
Tamaño de la muestra		-----Límites 3 Sigma para la media-----				
subg	del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.780156	15.957143	15.878947	16.977738	
2	7	14.780156	15.771429	15.878947	16.977738	
3	7	14.780156	16.442857	15.878947	16.977738	
4	7	14.780156	16.514286	15.878947	16.977738	
5	7	14.780156	16.071429	15.878947	16.977738	
6	7	14.780156	14.814286	15.878947	16.977738	
7	7	14.780156	15.914286	15.878947	16.977738	
8	7	14.780156	16.328571	15.878947	16.977738	
9	6	14.692117	15.216667	15.878947	17.065777	
10	7	14.780156	15.714286	15.878947	16.977738	
11	7	14.780156	15.828571	15.878947	16.977738	

Resumen del diagrama de medias y rangos para y					
-----Límites 3 Sigma para el rango-----					
subg	Límite inferior	Rango subgrupo	Rango Media	Límite superior	Límite sobrepasado
1	0.19840251	2.8000000	2.6206370	5.0428714	
2	0.19840251	3.0000000	2.6206370	5.0428714	
3	0.19840251	2.4000000	2.6206370	5.0428714	
4	0.19840251	1.1000000	2.6206370	5.0428714	
5	0.19840251	1.1000000	2.6206370	5.0428714	
6	0.19840251	2.5000000	2.6206370	5.0428714	
7	0.19840251	3.2000000	2.6206370	5.0428714	
8	0.19840251	1.9000000	2.6206370	5.0428714	
9	0.00000000	3.9000000	2.4559540	4.9213138	
10	0.19840251	4.1000000	2.6206370	5.0428714	
11	0.19840251	2.8000000	2.6206370	5.0428714	

Carta Xbar-R



Los resultados obtenidos son casi iguales excepto por algunas centésimas. En este caso solo se tiene un subgrupo de tamaño diferente, lo cual indica que para esta situación el método usual funciona bien.

Continuando con la exploración de las opciones en SAS, se puede fijar el tamaño muestral para el cálculo de los límites de control como se muestra abajo, usando la opción `LIMITN = 6` (tamaño de subgrupo nominal igual a 6).

Código SAS 2.4

```
options ls=80 ps=60 nodate nocenter nonumber;
options colors=(black) htitle=2 htext=1.6;
dm 'output;clear'; dm 'log;clear';
Data Calidad.tabla4_4M; set calidad.tabla4_4;
where y<>10.9; run;
TITLE 'Carta Xbar-R'; symbol v=dot;
PROC Shewhart data=calidad.tabla4_4M;
xrchart y*subg/limitn=6 alln nmarkers tableall;run;
```

32 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

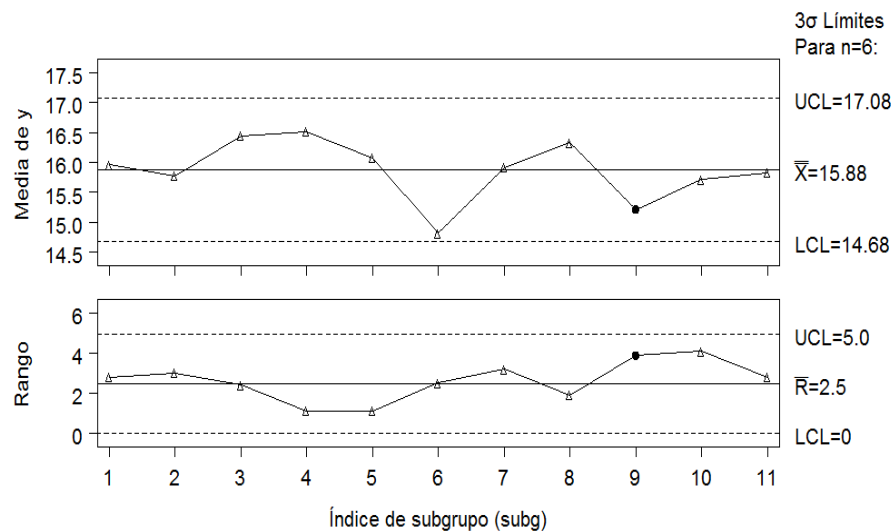
Salida SAS 2.4

Carta Xbar-R
Procedimiento SHEWHART

Resumen del diagrama de medias y rangos para y						
-----Límites 3 Sigma con n=6 para la media-----						
subg	Tamaño de la muestra del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.682463	15.957143	15.878947	17.075432	
2	7	14.682463	15.771429	15.878947	17.075432	
3	7	14.682463	16.442857	15.878947	17.075432	
4	7	14.682463	16.514286	15.878947	17.075432	
5	7	14.682463	16.071429	15.878947	17.075432	
6	7	14.682463	14.814286	15.878947	17.075432	
7	7	14.682463	15.914286	15.878947	17.075432	
8	7	14.682463	16.328571	15.878947	17.075432	
9	6	14.682463	15.216667	15.878947	17.075432	
10	7	14.682463	15.714286	15.878947	17.075432	
11	7	14.682463	15.828571	15.878947	17.075432	

Resumen del diagrama de medias y rangos para y						
-----Límites 3 Sigma con n=6 para el rango-----						
subg	Límite inferior	Rango subgrupo	Media subgrupo	Límite superior	Límite sobrepasado	
1	0	2.8000000	2.4759330	4.9613484		
2	0	3.0000000	2.4759330	4.9613484		
3	0	2.4000000	2.4759330	4.9613484		
4	0	1.1000000	2.4759330	4.9613484		
5	0	1.1000000	2.4759330	4.9613484		
6	0	2.5000000	2.4759330	4.9613484		
7	0	3.2000000	2.4759330	4.9613484		
8	0	1.9000000	2.4759330	4.9613484		
9	0	3.9000000	2.4759330	4.9613484		
10	0	4.1000000	2.4759330	4.9613484		
11	0	2.8000000	2.4759330	4.9613484		

Carta Xbar-R



Tamaños de subgrupo: ● n=6 Δ 6 < n ≤ 7

Para que salgan todos los subgrupos, independientemente del tamaño muestral, se da la opción ALLN, y con la opción NMARKERS se pide un símbolo especial para los subgrupos de tamaño muestral distinto al preespecificado, que en este caso es 6. Dada la estabilidad de los datos bajo análisis, los resultados son muy similares sin nada que destacar. Se puede pensar del gráfico de $\bar{X}$ que el subgrupo 6 tiene una media muy baja (a pesar de que no está fuera de límites), lo cual es confirmado por los ingenieros de calidad como causa asignable. Así que se resuelve eliminar el subgrupo 6 y recalcular los límites de control:

Código SAS 2.5

```
options ls=80 ps=60 nodate nocenter nonumber;
goptions colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
Data Calidad.tabla4_4M1;
set calidad.tabla4_4M1;
where subg<>6;
run;
TITLE 'Carta Xbar-R';
symbol v=dot;
PROC Shewhart data=calidad.tabla4_4M1;
xrchart y*subg/tableall;
run;
```

Salida SAS 2.5

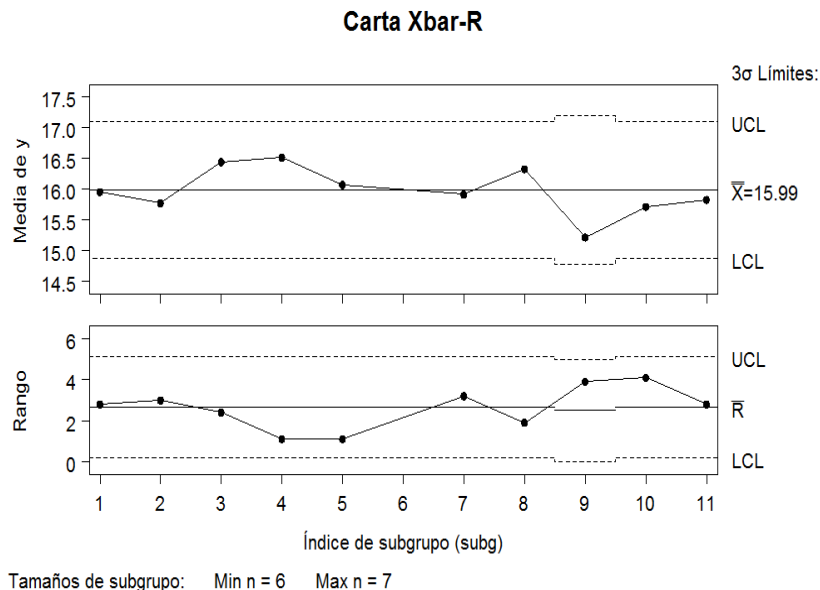
Carta Xbar-R
Procedimiento SHEWHART

Resumen del diagrama de medias y rangos para y						
-----Límites 3 Sigma para la media-----						
subg	Tamaño de la muestra del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.873275	15.957143	15.986957	17.100638	
2	7	14.873275	15.771429	15.986957	17.100638	
3	7	14.873275	16.442857	15.986957	17.100638	
4	7	14.873275	16.514286	15.986957	17.100638	
5	7	14.873275	16.071429	15.986957	17.100638	
7	7	14.873275	15.914286	15.986957	17.100638	
8	7	14.873275	16.328571	15.986957	17.100638	
9	6	14.784043	15.216667	15.986957	17.189870	
10	7	14.873275	15.714286	15.986957	17.100638	
11	7	14.873275	15.828571	15.986957	17.100638	

Resumen del diagrama de medias y rangos para y					
-----Límites 3 Sigma para el rango-----					
subg	Límite inferior	Rango subgrupo	Rango Media	Límite superior	Límite sobrepasado
1	0.20109122	2.8000000	2.6561513	5.1112114	
2	0.20109122	3.0000000	2.6561513	5.1112114	
3	0.20109122	2.4000000	2.6561513	5.1112114	
4	0.20109122	1.1000000	2.6561513	5.1112114	
5	0.20109122	1.1000000	2.6561513	5.1112114	
7	0.20109122	3.2000000	2.6561513	5.1112114	
8	0.20109122	1.9000000	2.6561513	5.1112114	
9	0.00000000	3.9000000	2.4892365	4.9880064	
10	0.20109122	4.1000000	2.6561513	5.1112114	
11	0.20109122	2.8000000	2.6561513	5.1112114	

34 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

Al observar los resultados, se aprecia una situación de mayor estabilidad, sobre todo en $\bar{X}$, lo cual hace suponer que se está en situación bajo control.



Para finalizar esta exploración de las cartas $\bar{X} - R$ usando SAS, se parte de los últimos datos sin incluir el subgrupo 6, y se calcularán las cartas $\bar{X} - R$ con límites probabilísticos usando la opción ALPHA = .002 en el programa que viene a continuación. Estos límites se calculan así:

Carta $\bar{X}$:

$$LCL = \bar{\bar{X}} - z_{\alpha/2} (\hat{\sigma}/n_i)$$

$$UCL = \bar{\bar{X}} + z_{\alpha/2} (\hat{\sigma}/n_i)$$

donde $z_{\alpha/2}$ es el percentil ($\alpha/2$) de la distribución normal estándar.

Carta R

$$LCL = D_{\alpha/2} \hat{\sigma}$$

$$UCL = D_{1-\alpha/2} \hat{\sigma}$$

donde $D_{\alpha/2}$ es el percentil ($\alpha/2$) de la distribución del rango de n observaciones independientes de una población normal con desviación estándar unitaria.

Código SAS 2.6

```
options ls=80 ps=60 nodate nocenter nonumber;
goptions colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
Data Calidad.tabla4_4M1;
set calidad.tabla4_4M;
where subg<>6; run;
TITLE 'Carta Xbar-R';
symbol v=dot;
PROC Shewhart data=calidad.tabla4_4M1;
xrchart y*subg/tableall alpha=.002;
run;
```

Salida SAS 2.6

Carta Xbar-R
Procedimiento SHEWHART

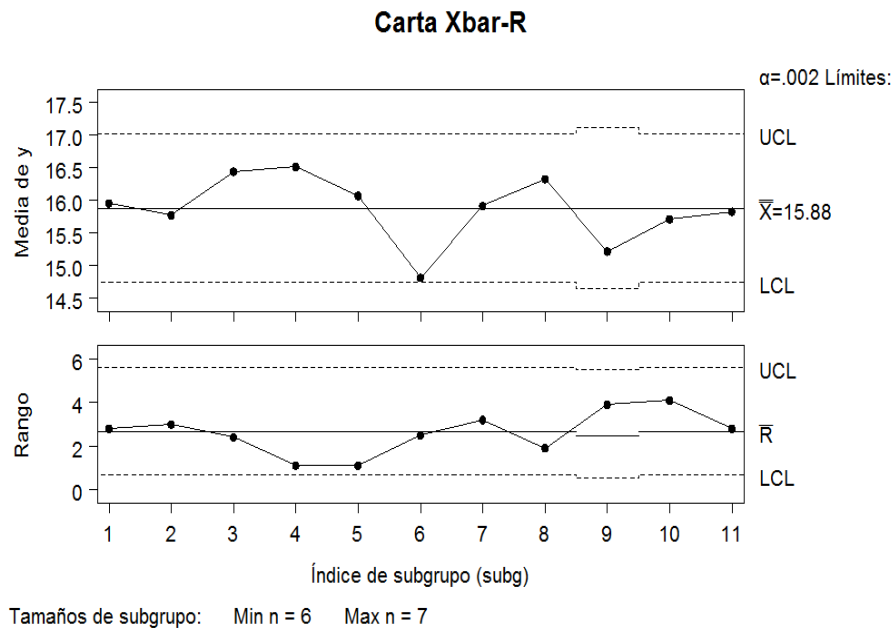
Resumen del diagrama de medias y rangos para y						
-----Límites Alpha=.002 para la media-----						
subg	Tamaño de la muestra del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.737900	15.957143	15.878947	17.019995	
2	7	14.737900	15.771429	15.878947	17.019995	
3	7	14.737900	16.442857	15.878947	17.019995	
4	7	14.737900	16.514286	15.878947	17.019995	
5	7	14.737900	16.071429	15.878947	17.019995	
6	7	14.737900	14.814286	15.878947	17.019995	
7	7	14.737900	15.914286	15.878947	17.019995	
8	7	14.737900	16.328571	15.878947	17.019995	
9	6	14.646475	15.216667	15.878947	17.111419	
10	7	14.737900	15.714286	15.878947	17.019995	
11	7	14.737900	15.828571	15.878947	17.019995	

Resumen del diagrama de medias y rangos para y					
-----Límites Alpha=.002 para el rango-----					
subg	Límite inferior	Rango subgrupo	Rango Media	Límite superior	Límite sobrepasado
1	0.67539455	2.8000000	2.6419557	5.5975441	
2	0.67539455	3.0000000	2.6419557	5.5975441	
3	0.67539455	2.4000000	2.6419557	5.5975441	
4	0.67539455	1.1000000	2.6419557	5.5975441	
5	0.67539455	1.1000000	2.6419557	5.5975441	
6	0.67539455	2.5000000	2.6419557	5.5975441	
7	0.67539455	3.2000000	2.6419557	5.5975441	
8	0.67539455	1.9000000	2.6419557	5.5975441	
9	0.52239763	3.9000000	2.4759330	5.4896705	
10	0.67539455	4.1000000	2.6419557	5.5975441	
11	0.67539455	2.8000000	2.6419557	5.5975441	

Los resultados no varían mucho con respecto al caso anterior. Se observa una ligera modificación en la carta R, donde se suben el UCL y el LCL, pero sin ningún resultado digno de comentar. En conclusión, para el ejemplo se considera que el proceso está bajo control en el caso donde se eliminó el subgrupo 6.

Con estos límites de control establecidos, se podría iniciar la Fase II, de monitoreo, con el objeto de mantener el proceso bajo control.

36 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS



A continuación se harán análisis similares con las cartas $\bar{X}-s$ usando el PROC SHEWHART de SAS y con el mismo ejemplo.

Se empieza el análisis con todos los datos. Más adelante está el programa con sus resultados; la gráfica superior es la carta $\bar{X}$ y la inferior es la carta s ; las tablas numéricas corresponden a los valores de cada carta.

Código SAS 2.7

```
options ls=80 ps=60 nodate nocenter nonumber;
goptions colors=(black) htitle=2 htext=1.6;
dm 'output;clear'; dm 'log;clear'; TITLE 'Carta Xbar-S';
symbol v=dot; PROC Shewhart data=sasuser.tabla4_4;
xschart y*subg/tableall; run;
```

Salida SAS 2.7

Carta Xbar-S
Procedimiento SHEWHART

Resumen del diagrama de medias y desviaciones típicas para y						
Tamaño de la muestra -----Límites 3 Sigma con n=7 para la media-----						
subg	del subgrupo	Límite inferior	Media subgrupo	Media promedio	Límite superior	Límite sobrepasado
1	7	14.594662	15.957143	15.814286	17.033910	
2	7	14.594662	15.771429	15.814286	17.033910	
3	7	14.594662	16.442857	15.814286	17.033910	
4	7	14.594662	16.514286	15.814286	17.033910	
5	7	14.594662	16.071429	15.814286	17.033910	
6	7	14.594662	14.814286	15.814286	17.033910	
7	7	14.594662	15.914286	15.814286	17.033910	
8	7	14.594662	16.328571	15.814286	17.033910	
9	7	14.594662	14.600000	15.814286	17.033910	

2.5. ANÁLISIS EN SAS USANDO PROC SHEWHART

37

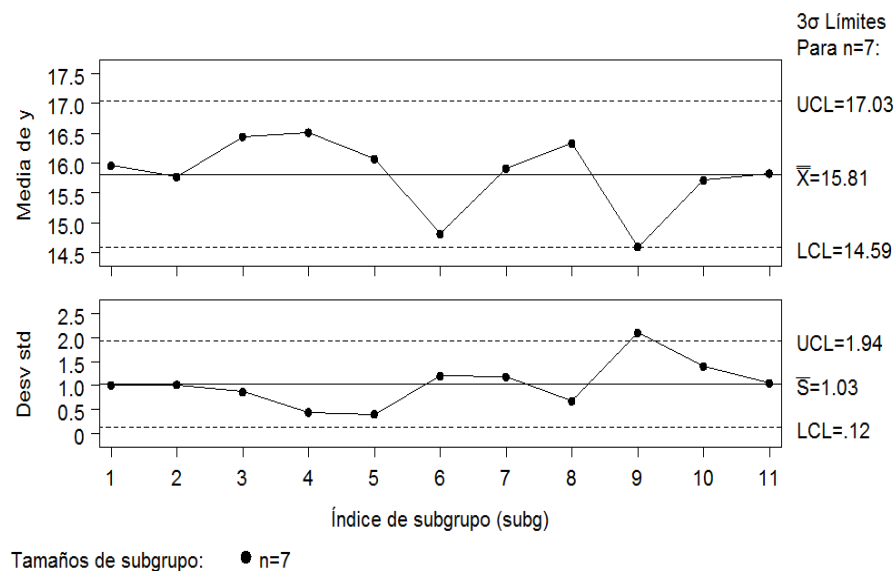
10	7	14.594662	15.714286	15.814286	17.033910
11	7	14.594662	15.828571	15.814286	17.033910

Resumen del diagrama de medias y desviaciones típicas para y

-----Límites 3 Sigma con n=7 para Dev std-----

subg	Límite inferior	Dev std subgrupo	Dev std media	Límite superior	Límite sobrepasado
1	0.12143966	1.0080627	1.0319041	1.9423685	
2	0.12143966	1.0160615	1.0319041	1.9423685	
3	0.12143966	0.8657504	1.0319041	1.9423685	
4	0.12143966	0.4413184	1.0319041	1.9423685	
5	0.12143966	0.4029652	1.0319041	1.9423685	
6	0.12143966	1.1978155	1.0319041	1.9423685	
7	0.12143966	1.1810004	1.0319041	1.9423685	
8	0.12143966	0.6799860	1.0319041	1.9423685	
9	0.12143966	2.1023796	1.0319041	1.9423685	Superior
10	0.12143966	1.4040757	1.0319041	1.9423685	
11	0.12143966	1.0515295	1.0319041	1.9423685	

Carta Xbar-S



Al observar las gráficas y compararlas con el análisis similar en las cartas $\bar{X}-R$ (páginas 29-31), se toma la misma decisión, esto es, eliminar el valor 10,9 de la lámina 9. A continuación está el programa SAS con los resultados para dichos cálculos.

38 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

Código SAS 2.8

```
options ls=80 ps=60 nodate nocenter nonumber;
goptions colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
Data sasuser.tabla4_4M;
set sasuser.tabla4_4;
where y<=10.9;
run;
TITLE 'Carta Xbar-S'; symbol v=dot;
PROC Shewhart data=sasuser.tabla4_4M;
xschart y*subg/tableall; run;
```

Salida SAS 2.8

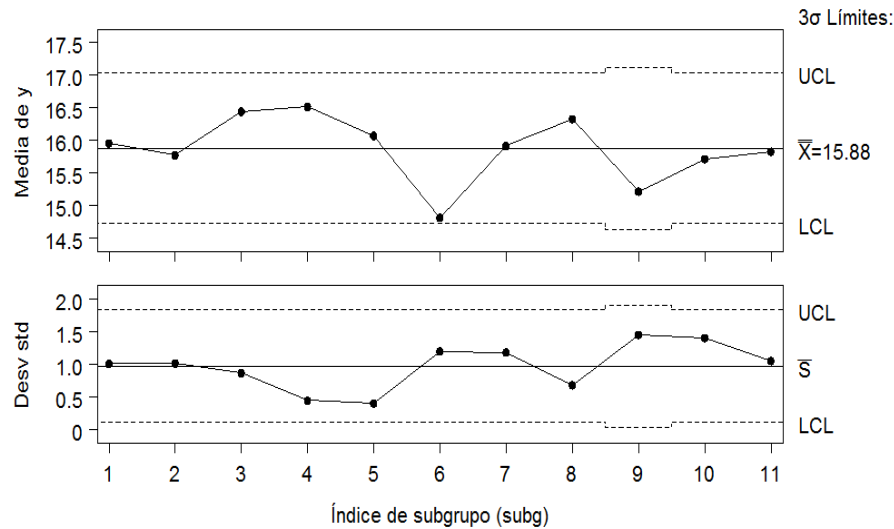
Carta Xbar-S
Procedimiento SHEWHART

Resumen del diagrama de medias y desviaciones típicas para y						
Tamaño de la muestra		-----Límites 3 Sigma para la media-----				
subg	del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.727869	15.957143	15.878947	17.030026	
2	7	14.727869	15.771429	15.878947	17.030026	
3	7	14.727869	16.442857	15.878947	17.030026	
4	7	14.727869	16.514286	15.878947	17.030026	
5	7	14.727869	16.071429	15.878947	17.030026	
6	7	14.727869	14.814286	15.878947	17.030026	
7	7	14.727869	15.914286	15.878947	17.030026	
8	7	14.727869	16.328571	15.878947	17.030026	
9	6	14.635641	15.216667	15.878947	17.122254	
10	7	14.727869	15.714286	15.878947	17.030026	
11	7	14.727869	15.828571	15.878947	17.030026	

Resumen del diagrama de medias y desviaciones típicas para y					
-----Límites 3 Sigma para Dev std-----					
subg	Límite inferior	Dev std subgrupo	Dev std media	Límite superior	Límite sobrepasado
1	0.11461448	1.0080627	0.97390869	1.8332029	
2	0.11461448	1.0160615	0.97390869	1.8332029	
3	0.11461448	0.8657504	0.97390869	1.8332029	
4	0.11461448	0.4413184	0.97390869	1.8332029	
5	0.11461448	0.4029652	0.97390869	1.8332029	
6	0.11461448	1.1978155	0.97390869	1.8332029	
7	0.11461448	1.1810004	0.97390869	1.8332029	
8	0.11461448	0.6799860	0.97390869	1.8332029	
9	0.02932946	1.4524692	0.96595401	1.9025785	
10	0.11461448	1.4040757	0.97390869	1.8332029	
11	0.11461448	1.0515295	0.97390869	1.8332029	

Se observa que los límites de control para el subgrupo 9 son diferentes debido a que el tamaño de este subgrupo es ahora de 6. Para ello SAS, como en el caso de las cartas $\bar{X} - R$, usa fórmulas que tienen en cuenta el tamaño del subgrupo (ver SAS/QC, 1999, p. 1574). Del gráfico de $\bar{X}$ se observa que el subgrupo 6 tiene una media muy baja (a pesar de que no está fuera de límites), lo cual es confirmado por los ingenieros de calidad como causa asignable. Así que se resuelve eliminar el subgrupo 6 y recalcular los límites de control.

Carta Xbar-S



Tamaños de subgrupo: Min n = 6 Max n = 7

Código SAS 2.9

```
options ls=80 ps=60 nodate nocenter nonumber;
options colors=(black) htitle=2 htext=1.6;
dm 'output;clear';
dm 'log;clear';
Data sasuser.tabla4_4M1;
set sasuser.tabla4_4M;
where subg<>6;
run;
TITLE 'Carta Xbar-S';
symbol v=dot;
PROC Shewhart data=sasuser.tabla4_4M1;
xschart y*subg/tableall; run;
```

Salida SAS 2.9

Carta Xbar-S
Procedimiento SHEWHART

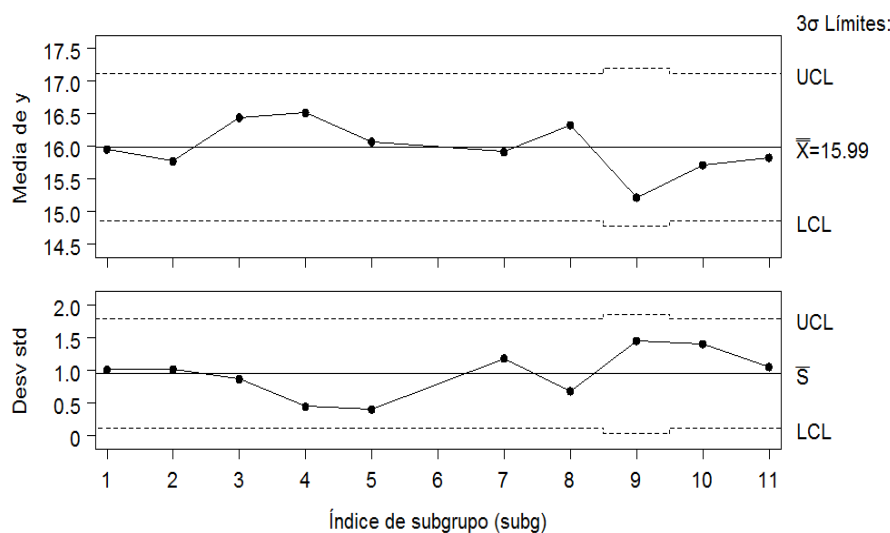
Resumen del diagrama de medias y desviaciones típicas para y						
Tamaño de la muestra		-----Límites 3 Sigma para la media-----				
subg	del subgrupo	Límite inferior	Media subgrupo	Media Promedio	Límite superior	Límite sobrepasado
1	7	14.862342	15.957143	15.986957	17.111571	
2	7	14.862342	15.771429	15.986957	17.111571	
3	7	14.862342	16.442857	15.986957	17.111571	
4	7	14.862342	16.514286	15.986957	17.111571	
5	7	14.862342	16.071429	15.986957	17.111571	
7	7	14.862342	15.914286	15.986957	17.111571	
8	7	14.862342	16.328571	15.986957	17.111571	
9	6	14.772234	15.216667	15.986957	17.201679	

40 CAPÍTULO 2. CARTAS DE CONTROL PARA MEDICIONES CON SUBGRUPOS

10	7	14.862342	15.714286	15.986957	17.111571
11	7	14.862342	15.828571	15.986957	17.111571

Resumen del diagrama de medias y desviaciones típicas para y					
-----Límites 3 Sigma para Dev std-----					
subg	Límite inferior	Dev std subgrupo	Dev std media	Límite superior	Límite sobrepasado
1	0.11197943	1.0080627	0.95151801	1.7910566	
2	0.11197943	1.0160615	0.95151801	1.7910566	
3	0.11197943	0.8657504	0.95151801	1.7910566	
4	0.11197943	0.4413184	0.95151801	1.7910566	
5	0.11197943	0.4029652	0.95151801	1.7910566	
7	0.11197943	1.1810004	0.95151801	1.7910566	
8	0.11197943	0.6799860	0.95151801	1.7910566	
9	0.02865516	1.4524692	0.94374621	1.8588373	
10	0.11197943	1.4040757	0.95151801	1.7910566	
11	0.11197943	1.0515295	0.95151801	1.7910566	

Carta Xbar-S



Tamaños de subgrupo: Min n = 6 Max n = 7

Al observar los resultados, se aprecia una situación de mayor estabilidad, sobre todo en $\bar{X}$, lo cual hace suponer que se está en situación bajo control.

En resumen, los resultados con las cartas $\bar{X}-R$ y $\bar{X}-s$ son idénticos. Esto era de esperarse pues el tamaño del subgrupo es menor que 10, en cuyo caso R y s se comportan de manera similar (ver página 28).

2.6 Longitud promedio de corridas

Se refiere a la sigla en inglés *Average Run Length* que se traduce como longitud promedio de corridas, concepto que se explicará a continuación (ver Vargas, 2001).

Una vez construidos los límites de una carta de control, al número de puntos graficados en la carta hasta que aparezca una señal fuera de control se le conoce por longitud de corrida (RL).

El valor de RL cambia de ensayo a ensayo debido a la variabilidad aleatoria. El valor esperado de la variable aleatoria RL, con base en un número grande de ensayos, se denomina ARL. El ARL es criterio importante para comparar la eficiencia de varias cartas de control.

Considérese una carta $\bar{X}$ con límites de control UCL y LCL. Supóngase que se toman grupos de n observaciones en cada tiempo y se grafica $\bar{X}_i$ en la carta. Las observaciones son independientes e idénticamente distribuidas como una $N(\mu, \sigma^2)$; sea

$$p = P[\bar{X}_i > UCL] + P[\bar{X}_i < LCL]$$

la probabilidad de que un punto esté “fuera de control”, y sea C la variable aleatoria definida por C = número de puntos graficados en la carta antes de obtener el primero fuera de control.

Claramente C sigue una distribución geométrica con parámetro p . Es decir

$$P[C = x] = p(1 - p)^{x-1}, \quad x = 1, 2, 3, \dots$$

luego

$$E[C] = \frac{1}{p} \quad y \quad Var[C] = \frac{1-p}{p^2}$$

Por tanto,

$$ARL = E[C] = \frac{1}{p}$$

Es deseable para el caso “en control” que el ARL sea razonablemente grande, de tal manera que las falsas alarmas raramente ocurran. Bajo normalidad en el caso de una carta $\bar{X}$ con límites 3σ , $p = 0,0027$.

Así que

$$ARL = E[C] \approx 370,37$$

Esto es, se espera cada 370 subgrupos tener una falsa alarma, es decir un subgrupo fuera de control debido a causas puramente aleatorias.

El ARL se puede calcular también mediante cadenas de Markov. El lector interesado puede encontrar detalles en Vargas (2001).

Ahora bien, cuando un parámetro cambia en una cantidad que tiene consecuencias, se quiere detectar el cambio tan rápido como sea posible. En consecuencia, el ARL para cambio de parámetro debe ser pequeño. Infortunadamente, las cartas tipo Shewhart no tienen buenas propiedades ARL.

Por ejemplo, supóngase un aumento de un $\sigma_{\bar{x}}$ en la media cuando se usa una carta $\bar{X}$. Asímanse parámetros conocidos y determínese la probabilidad de que un punto caiga fuera de los límites de control:

Para

$$UCL : Z = [\mu + 3\sigma_{\bar{x}} - (\mu + \sigma_{\bar{x}})] / \sigma_{\bar{x}} = 2$$

Para

$$LCL : Z = [\mu - 3\sigma_{\bar{x}} - (\mu + \sigma_{\bar{x}})] / \sigma_{\bar{x}} = -4$$

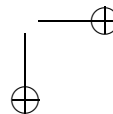
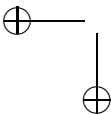
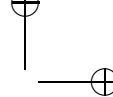
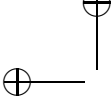
Entonces

$$ARL = \frac{1}{P[Z > 2] + P[Z < -4]} = 43,89$$

Este es un número alto para una rápida detección del cambio en la media. Evidentemente, un buen sistema de control debería tener una tasa de falsa alarma grande para evitar sobrecontrol (en el caso normal fue de 370; a este ARL se le suele denotar por L_0 y se le denomina ARL en control; ver Bissell, 1994), y un ARL de cambio de parámetro corto (a este último se le suele denotar, por ejemplo, $L_1 = 10$ para $\Delta = 1.6\sigma_{\bar{x}}$, donde Δ se usa para indicar el cambio en el nivel de la media).

Otro criterio más general para medir desempeño en un carta de control es el ATS (*Average Time to Signal*). Si las muestras se toman en

tiempos iguales, por ejemplo cada hora, $ATS = ARL \times 1$. En este caso no es necesario considerar ATS , pues se asume que las muestras se tomaron en tiempos separados iguales.



CAPÍTULO 3

Cartas de control para observaciones individuales

En algunas situaciones es difícil agrupar las observaciones de ciertos procesos. Por ejemplo, si un ítem es producido cada 30 minutos, tomaría 2,5 horas formar un subgrupo de 5 ítems, y para entonces el proceso podría haberse salido de control; este es el caso de una línea de ensamblaje con una tasa baja de producción. Otros ejemplos de variables que se deberían medir individualmente, debido a su naturaleza, son la temperatura, la presión y los datos de contabilidad de empresas de servicios. En el caso de la temperatura y la presión, cinco mediciones rápidas serían prácticamente la misma, lo cual no agrega ninguna información.

3.1 Cartas para observaciones individuales

Las cartas X basadas en observaciones individuales no son tan sensibles como las cartas $\bar{X}$ para detectar ‘desviaciones’ en la media. Supóngase que hay una desviación de la media de $a\sigma$, donde $a > 0$. Los límites de control para una carta X son $\mu \pm 3\sigma$. Asúmase que los parámetros son conocidos; entonces:

$$P(1 \text{ observación individual} > UCL) = P(X > \mu + 3\sigma),$$

dado que la media real es igual a $\mu + a\sigma$. Ahora bien, para la carta $\bar{X}$ se necesita determinar $P(\bar{X} > \mu + 3\sigma_{\bar{X}})$. Así que los dos valores de z son:

$$z = \frac{\mu + 3\sigma - (\mu + a\sigma)}{\sigma} = 3 - a$$

y

$$z = \frac{\mu + 3\sigma_{\bar{X}} - (\mu + a\sigma)}{\sigma_{\bar{X}}} = \frac{\mu + 3\sigma/\sqrt{n} - (\mu + a\sigma)}{\sigma/\sqrt{n}} = 3 - a\sqrt{n}$$

Claramente, $P(Z > 3 - a\sqrt{n}) > P(Z > 3 - a)$. Por tanto, la probabilidad de observar un ‘subgrupo promedio’ fuera del UCL cuando se tiene una desviación en la media $a\sigma$, es mayor que la de observar un solo ítem.

Este resultado se podría argumentar desde la teoría estadística, sobre el hecho de que la potencia de una prueba de hipótesis se incrementa cuando el tamaño de muestra se incrementa (en este punto se estarían comparando muestras de tamaño uno, con tamaños de muestra de $n > 1$, suponiendo que las medidas de los subgrupos fueron hechas en el mismo punto en el tiempo).

En general, las cartas $\bar{X}$ son mejores que las cartas X . Sin embargo, se recomienda el uso conjunto de los dos tipos de cartas, pues podría revelar, en ciertos casos, algún tipo de error en la medición de las observaciones. Por ejemplo, si un dato es excesivamente grande, una carta X lo detectaría, y se podría identificar si hay o no error de medición. Con el mismo propósito, se podrían emplear las cartas X junto con las cartas R , s o s^2 .

3.1.1 Límites de control para la carta X

Los límites de control para una carta X cuando los parámetros son desconocidos están dados por:

$$\hat{\mu} \pm 3\hat{\sigma}$$

Según Ryan (2000), en la Fase I se recomienda estimar los parámetros con datos históricos de al menos 100 observaciones, en tanto que Quesenberry (1993) recomienda un tamaño de al menos 300 observaciones.

Puesto que no hay subgrupos, obviamente no se pueden emplear los rangos por subgrupos para la estimación de σ . El procedimiento más usado es crear rangos tomando el valor absoluto de las diferencias de las observaciones sucesivas (segunda menos la primera, tercera menos la segunda, y así sucesivamente). El promedio de estos ‘rangos móviles’ (MR) de tamaño 2 se usa de la misma forma que $\bar{R}$ al estimar σ para una carta $\bar{X}$, es decir:

$$\hat{\sigma} = \frac{\overline{MR}}{d_2}$$

donde $\overline{MR}$ denota el promedio de los rangos móviles y d_2 es la constante tabulada usada para hacer al estimador insesgado (ver Ryan, 2000, tabla E).

Esta aproximación es válida para la Fase I, pues el proceso por monitorear está probablemente fuera de control, en cuyo caso el estimador basado en rangos móviles es más robusto (por ejemplo, insensible) que el estimador basado en la desviación estándar, relativo al tipo de condiciones fuera de control que podrían haber ocurrido en la Fase I (Ryan, 2000, p. 135).

Cuando se usa una carta X , se está asumiendo que los datos son aproximadamente normales, y se sabe que el estimador de sigma basado en los rangos móviles es más robusto que el basado en la desviación estándar muestral. Ahora bien, si la distribución es normal, Cryer & Ryan (1990) muestran que es mucho mejor estimar a sigma como $\hat{\sigma} = s/c_4$ donde c_4 es la constante tabulada que hace el estimador insesgado (ver Ryan, 2000, tabla E). Específicamente,

$$Var(\overline{MR}/d_2)/Var(s/c_4) = 1,65$$

independiente del número de observaciones históricas usadas en la determinación de los límites que luego se usarán en la Fase II para monitorear el proceso. Esto quiere decir que si μ fuera conocida y se usara la aproximación de rangos móviles en lugar de la estimación de σ usando la desviación estándar, se requeriría, grosso modo, 65 % más de observaciones para que la carta X tenga el ARL que tendría si se usara el estimador basado en σ .

3.1.2 Supuestos de la carta $\bar{X}$

Los supuestos para la carta $\bar{X}$ son los mismos que para la carta $\bar{X}$: normalidad e independencia. El supuesto de normalidad es más importante cuando se grafican las observaciones individuales, puesto que no se puede aplicar el teorema del límite central. Incluso una pequeña desviación de la normalidad puede hacer que el ARL en control sea mucho menor que el valor bajo la teoría normal. Por ejemplo, si se comparan dos cartas de control $\bar{X}$, una en que la variable sigue una distribución chi cuadrado con 30 gl, con otra en que la variable sigue una distribución normal, se tiene que el ARL bajo control de la segunda es 370,37 mientras que el de la primera disminuye a menos de la mitad (ARL = 179,04).

Relativo a la autocorrelación, puesto que en un proceso $AR(1)$, $E(s/c_4) \rightarrow \sigma$ cuando $n \rightarrow \infty$, no habría problema al usar la desviación estándar muestral para estimar σ cuando se tienen muestras grandes. En conclusión, el problema de normalidad es más serio que el de autocorrelación cuando se grafican observaciones individuales. La autocorrelación puede ser compensada ajustando los límites de control al usar $\hat{\sigma} = s/c_4$ y al tener el estimador basado en un gran número de observaciones.

3.1.3 Ejemplo ilustrativo usando SAS

Para este ejemplo se tomó una muestra de 50 números aleatorios, provenientes de una distribución $N(\mu = 25, \sigma^2 = 9)$, tal y como se muestra en la tabla 3.1. Se asumirá que estos números aleatorios constituyen 50 observaciones consecutivas de un proceso de manufactura.

TABLA 3.1 Números aleatorios de una distribución $N(\mu = 25, \sigma^2 = 9)$.

28,30	17,89	26,45	24,69	24,18	23,64	28,38	26,83	24,35	24,54
26,92	22,71	23,75	16,79	21,35	30,53	24,35	24,26	24,00	27,09
26,68	23,80	27,60	25,66	24,30	25,87	30,80	25,74	30,70	29,54
24,26	25,54	29,16	21,27	27,11	20,03	24,85	25,86	21,27	21,50
21,58	23,58	27,03	26,43	23,89	22,24	24,14	24,28	28,01	24,33

Aquí se sabe que los datos están normalmente distribuidos, así que una prueba de normalidad daría como resultado la aceptación de la hipótesis nula. Sin embargo, para chequear normalidad, se presenta el programa en SAS a manera de ilustración y para efectos de completar el análisis.

Código SAS 3.1

```
TITLE 'Q - Q Plot';
TITLE2 'Observaciones de una Normal(25,9)';
symbol v=dot c=black h=0.7;
proc univariate data=tabla5_1 normaltest;
var x;
probplot x / normal (mu=est sigma=est)
square name='MyPlot';
inset mean std;
run;
```

Salida SAS 3.1

Test para normalidad

Test	-Estadístico--		-----P-valor-----	
Shapiro-Wilk	W	0.972411	Pr < W	0.2893
Kolmogorov-Smirnov	D	0.121933	Pr > D	0.0627
Cramer-von Mises	W-Sq	0.087783	Pr > W-Sq	0.1638
Anderson-Darling	A-Sq	0.488723	Pr > A-Sq	0.2209

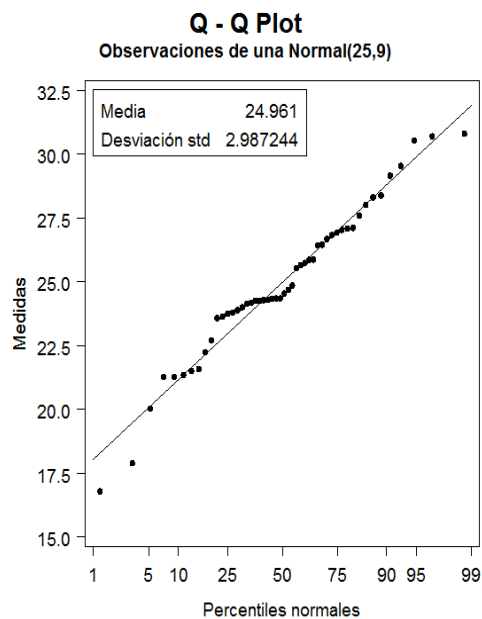


FIGURA 3.1 Gráfico de probabilidad normal para los datos de la tabla 3.1.

Obsérvese que en todos los test se acepta la normalidad, pues el valor-p, para cada uno de ellos, es mayor que 0,05.

50 CAPÍTULO 3. CARTAS DE CONTROL PARA OBSERVACIONES INDIVIDUALES

Para efectos de ilustración, σ (que se sabe toma un valor de 3) será estimada usando s y $\overline{MR}/d_2$ para un rango móvil de tamaño dos. Se puede ver que $s/c_4 = 2,9872/0,9949 = 3,0025$ y $\overline{MR}/d_2 = 3,1484/1,128 = 2,7911$. Así, los dos estimadores difieren, aun cuando las observaciones son números aleatorios distribuidos normalmente. Sin embargo, como se mencionó en la sección 3.1.1, el estimador $\overline{MR}/d_2$ se prefiere en la Fase I.

Ahora se analiza la construcción de estas cartas de control en SAS.

En SAS se pueden manejar las observaciones individuales con el argumento `irchart` del PROC SHEWHART. Aquí la desviación estándar es estimada como $\overline{MR}/d_2$. Se mostrarán las cartas X y de rangos móviles (denotadas en las gráficas como cartas $X - MR$). A continuación aparece el programa con sus resultados. La gráfica superior es la carta X y la inferior es el MR . Las tablas numéricas corresponden a los valores de cada carta.

Código SAS 3.2

```
options ls=80 ps=60 nodate nocenter nonumber;
options colors=(black) htitle=2 htext=1.6;
data tabla5_1;
inputx @@;
obs = _N_;
label x = "medidas"
obs = ''unidad";
cards;
28.30 23.64 26.92 30.53 26.68 25.87 24.26 20.03 21.58 22.24
17.89 28.38 22.71 24.35 23.80 30.80 25.54 24.85 23.58 24.14
26.45 26.83 23.75 24.26 27.60 25.74 29.16 25.86 27.03 24.28
24.69 24.35 16.79 24.00 25.66 30.70 21.27 21.27 26.43 28.01
24.18 24.54 21.35 27.09 24.30 29.54 27.11 21.50 23.89 24.33
run;
proc print data=tabla5_1;
run;
TITLE 'Carta de observaciones individuales y de Rangos Moviles';
TITLE2 'Observaciones de una Normal(25,9)';
symbol v=dot c=black;
proc shewhart data=tabla5_1;
irchart x*obs/tableall;
run;
```

Los resultados arrojados por el anterior programa son los siguientes:

Salida SAS 3.2

Carta de observaciones individuales y de rangos Móviles
Observaciones de una Normal(25,9)

```

Resumen del diagrama de medidas individuales para x
-----Límites 3 Sigma con n=2 for x-----
obs      Límite inferior      x      Media      Límite superior      Límite
1      16.590496      28.300000      24.961000      33.331504
2      16.590496      23.640000      24.961000      33.331504
3      16.590496      26.920000      24.961000      33.331504
4      16.590496      30.530000      24.961000      33.331504
5      16.590496      26.680000      24.961000      33.331504
.
.
46      16.590496      29.540000      24.961000      33.331504
47      16.590496      27.110000      24.961000      33.331504
48      16.590496      21.500000      24.961000      33.331504
49      16.590496      23.890000      24.961000      33.331504
50      16.590496      24.330000      24.961000      33.331504

Resumen del diagrama de medidas individuales para x
-----Límites 3 Sigma con n=2 para rango móvil-----
obs      Límite inferior      Rango      Rango      Límite superior      Límite
1      0      móvil      de la media      10.284242
2      0      4.660000      3.1483673      10.284242
3      0      3.280000      3.1483673      10.284242
4      0      3.610000      3.1483673      10.284242
5      0      3.850000      3.1483673      10.284242
.
.
12      0      10.490000      3.1483673      10.284242      Superior
.
.
46      0      5.240000      3.1483673      10.284242
47      0      2.430000      3.1483673      10.284242
48      0      5.610000      3.1483673      10.284242
49      0      2.390000      3.1483673      10.284242
50      0      0.440000      3.1483673      10.284242

```

Las cartas MR miden la variación entre diferencias sucesivas (ob-
sérvese en la carta X que en la parte superior derecha, segunda línea,
aparece $n = 2$, que se refiere a diferencias sucesivas de orden 2); para este
ejemplo se observa cómo la diferencia 12 da mayor que el promedio, como
se muestra en la figura 3.2 y en la salida de SAS 3.3 correspondiente a
este valor, donde aparece en la última columna la leyenda ‘Superior’. Al
examinar los datos, se aprecia que no hay ningún *outlier* y que, en la
carta X , esta observación queda dentro de los límites de control. Ahora
bien, a manera de ejercicio (en la práctica puede haber razones que ex-
pliquen el bajo valor de este dato), se eliminarán las observaciones 11 y
33, cuyos valores (17,89 y 16,79, respectivamente) son las observaciones
más pequeñas con respecto al promedio, y se recalcularán los límites de
control para ambas cartas de control, como se ilustra a continuación.

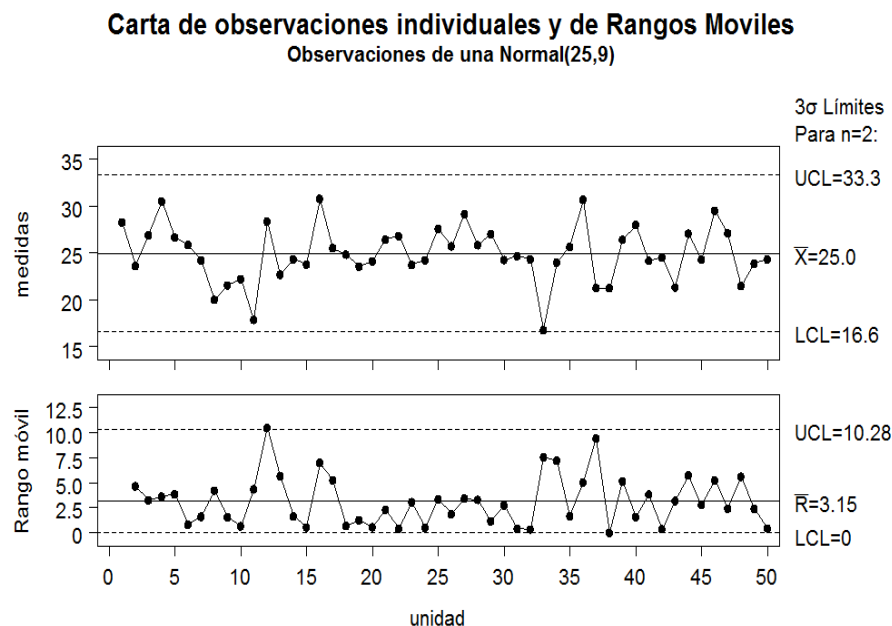


FIGURA 3.2 Carta $\bar{X}$ y de rangos móviles para observaciones individuales. σ estimado como $\overline{MR}/d_2$.

Código SAS 3.3

```
data tabla5_1B;
set tabla5_1; where x<>17.89 and x<>16.79;
run; TITLE 'Carta X-MR'; symbol v=dot c=black;
proc shewhart data=tabla5_1B; irchart x*obs/tableall; run;
```

Salida SAS 3.3

Resumen del diagrama de medidas individuales para x
-----Límites 3 Sigma con n=2 para x-----

obs	Límite inferior	x	Media	Límite superior	Límite sobrepasado
1	17.859691	28.300000	25.278542	32.697392	
2	17.859691	23.640000	25.278542	32.697392	
3	17.859691	26.920000	25.278542	32.697392	
4	17.859691	30.530000	25.278542	32.697392	
5	17.859691	26.680000	25.278542	32.697392	
.	.	.	.	.	
.	.	.	.	.	
44	17.859691	27.090000	25.278542	32.697392	
45	17.859691	24.300000	25.278542	32.697392	
46	17.859691	29.540000	25.278542	32.697392	
47	17.859691	27.110000	25.278542	32.697392	
48	17.859691	21.500000	25.278542	32.697392	
49	17.859691	23.890000	25.278542	32.697392	
50	17.859691	24.330000	25.278542	32.697392	

3.1. CARTAS PARA OBSERVACIONES INDIVIDUALES

53

Resumen del diagrama de medidas individuales para x
-----Límites 3 Sigma con n=2 para rango móvil-----

obs	Límite inferior	Rango móvil de la media	Límite superior	Límite sobrepasado
1	0	2.7904255	9.1150141	
2	0	4.6600000	9.1150141	
3	0	3.2800000	9.1150141	
4	0	3.6100000	9.1150141	
5	0	3.8500000	9.1150141	
.	.	.	.	.
.	.	.	.	.
36	0	5.0400000	9.1150141	
37	0	9.4300000	9.1150141	Superior
.	.	.	.	.
.	.	.	.	.
45	0	2.7900000	9.1150141	
46	0	5.2400000	9.1150141	
47	0	2.4300000	9.1150141	
48	0	5.6100000	9.1150141	
49	0	2.3900000	9.1150141	
50	0	0.4400000	9.1150141	

Carta X-MR

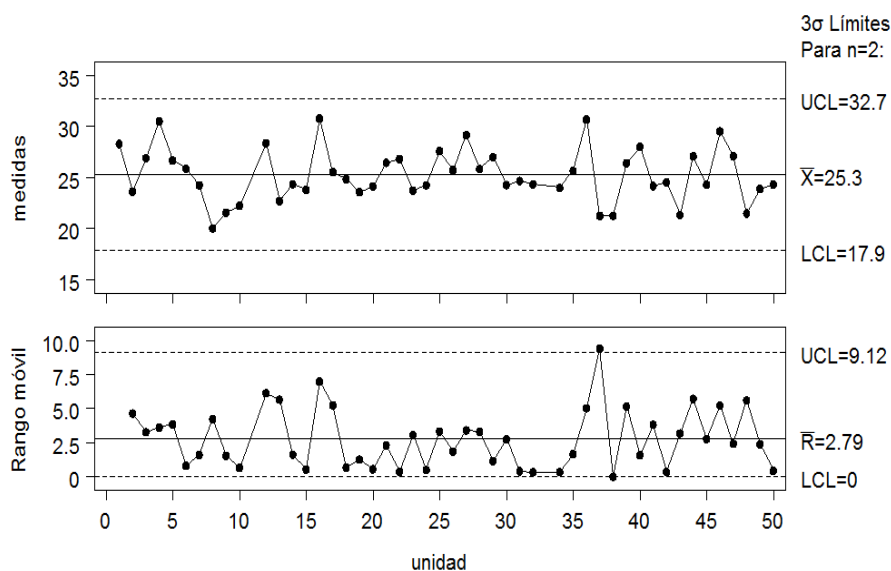


FIGURA 3.3 Carta X-MR para datos de la tabla 3.1, sin incluir las observaciones 11 y 33.

En la figura 3.3 se muestra un mejor comportamiento de los datos dentro de los límites de control, para la carta $\bar{X}$. Sin embargo, sigue habiendo problemas con la carta de rangos móviles (MR). A manera de ejercicio, suponga que los ingenieros examinan el dato 36, con valor 30,70,

54 CAPÍTULO 3. CARTAS DE CONTROL PARA OBSERVACIONES INDIVIDUALES

y que recomiendan sacarlo por problemas técnicos. En esta ocasión, los gráficos y las tablas numéricas son:

Código SAS 3.4

```
data tabla5_1B;
set tabla5_1;
where x<>17.89 and x<>16.79 and x<>30.70;
run; TITLE 'Carta X-MR';
symbol v=dot c=black;
proc shewhart data=tabla5_1B;
irchart x*obs/tableall;
run;
```

Salida SAS 3.4

```
Resumen del diagrama de medidas individuales para x
-----Límites 3 Sigma con n=2 para x-----
obs      Límite inferior      x      Media      Límite superior      Límite
1      18.165659      28.300000      25.163191      32.160724      sobrepasado
2      18.165659      23.640000      25.163191      32.160724
3      18.165659      26.920000      25.163191      32.160724
4      18.165659      30.530000      25.163191      32.160724
5      18.165659      26.680000      25.163191      32.160724
.      .      .      .      .
.      .      .      .      .
.      .      .      .      .
45      18.165659      24.300000      25.163191      32.160724
46      18.165659      29.540000      25.163191      32.160724
47      18.165659      27.110000      25.163191      32.160724
48      18.165659      21.500000      25.163191      32.160724
49      18.165659      23.890000      25.163191      32.160724
50      18.165659      24.330000      25.163191      32.160724
```

```
Resumen del diagrama de medidas individuales para x
-----Límites 3 Sigma con n=2 para rango móvil-----
obs      Límite inferior      Rango      Rango      Límite      Límite
1      0      .      móvil      de la media      superior      sobrepasado
2      0      4.6600000      2.6319565      8.5973700
3      0      3.2800000      2.6319565      8.5973700
4      0      3.6100000      2.6319565      8.5973700
5      0      3.8500000      2.6319565      8.5973700
.      .      .      .      .
.      .      .      .      .
.      .      .      .      .
.      .      .      .      .
45      0      2.7900000      2.6319565      8.5973700
46      0      5.2400000      2.6319565      8.5973700
47      0      2.4300000      2.6319565      8.5973700
48      0      5.6100000      2.6319565      8.5973700
49      0      2.3900000      2.6319565      8.5973700
50      0      0.4400000      2.6319565      8.5973700
```

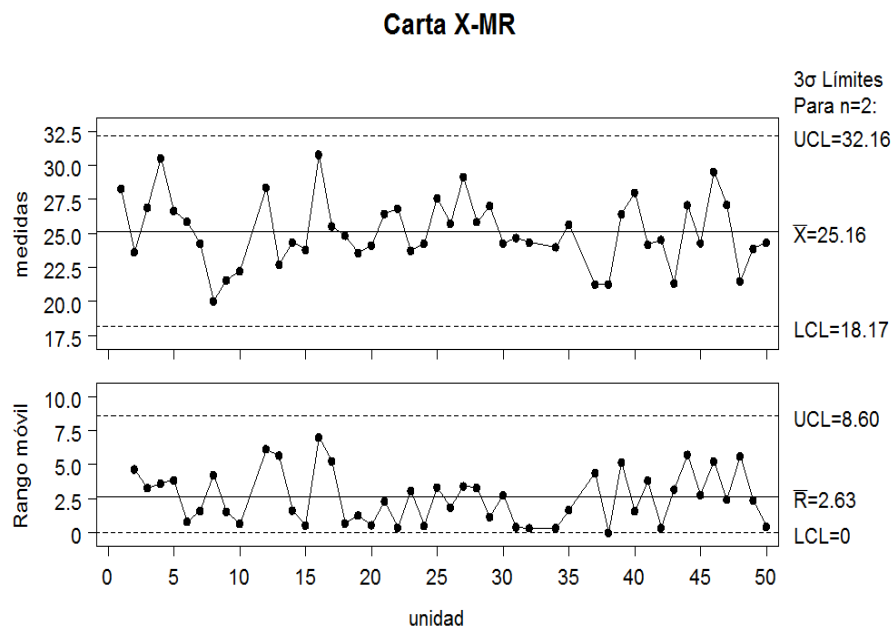


FIGURA 3.4 Carta X-MR para datos de la tabla 3.1, sin incluir las observaciones 11, 33 y 36.

Finalmente se obtienen cartas X y MR en control, que servirán como referente histórico para la Fase II. Los límites de control para las cartas X y MR se calculan en SAS como

- Cartas de medidas individuales (X)

$$LCL = \bar{X} - k\hat{\sigma}$$

$$UCL = \bar{X} + k\hat{\sigma}$$

- Cartas de rangos móviles (MR)

$$LCL = \max(d_2(n)\hat{\sigma} - kd_3(n)\hat{\sigma}, 0)$$

$$UCL = d_2(n)\hat{\sigma} + kd_3(n)\hat{\sigma}$$

donde

k : es un múltiplo de los errores estándar. Por defecto, los límites son calculados con $k = 3$ (estos son referidos como los límites 3σ). Este parámetro puede ser especificado con la variable `_SIGMAS_` en `LIMITS=`, como se muestra abajo en el programa SAS, con $k = 3$ (se deja como ejercicio al lector explorar con valores de k diferentes).

$d_2(n)$: el valor esperado de rango de n variables normales independientemente distribuidos con desviación estándar uno.

$d_3(n)$: el error estándar del rango de n observaciones independientes de una población normal con desviación estándar uno.

Ahora, si se quiere utilizar s/c_4 como estimación de σ , es necesario crear por aparte un DATA en el cual se especifique el valor de σ por emplear, y luego, con la opción `LIMITS` del `PROC SHEWHART`, se indica qué valores deberán ser empleados para el cálculo de los límites, es decir, toma un conjunto de datos diferentes para redefinir los límites, como sigue:

Código SAS 3.5

```
proc means data=tabla5_1 noprint;
var x;
output out=estad mean=media std=desv;
run;
proc print data=estad;
run;
```

El anterior procedimiento de SAS arroja los valores de la media y la desviación estándar, que serán utilizados para redefinir los límites de control, empleando s/c_4 .

Salida SAS 3.5

Obs	_TYPE_	_FREQ_	media	desv
1	0	50	24.961	2.98724

Código SAS 3.6

```
data limites;
length _var_ _subgrp_ _type_8;
_var_='x';
_subgrp_='obs';
_limitn_=2; /*este es el n=2 de las diferencias que se
podrían cambiar*/
_type_='standard';
_mean_=24.9604; /*esta es la media*/
_stddev_=2.98727/0.9949; /*esta es s/c4=desv/c4, con c4 tomada
de las tablas*/
_sigmas_=3;
run;
title 'Carta para observaciones individuales y para Rangos
Móviles';
title2 'Observaciones de una Normal(25,9)';
title3 'sigma estimado por s/c4';
symbol v=dot c=black;
proc shewhart data=tabla5_1 limits=limites;
irchart x*obs/tableall;
run;
```

Los resultados gráficos y las tablas numéricas de este último PROC se presentan a continuación.

Salida SAS 3.6

Carta para observaciones individuales y para rangos móviles
Observaciones de una Normal(25,9)
Sigma estimado por s/c4

Procedimiento SHEWHART

Resumen del diagrama de medidas individuales para x					
-----Límites 3 Sigma con n=2 para x-----					
obs	Límite inferior	x	Media	Límite superior	Límite sobrepasado
1	15.952650	28.300000	24.960400	33.968150	
2	15.952650	23.640000	24.960400	33.968150	
3	15.952650	26.920000	24.960400	33.968150	
4	15.952650	30.530000	24.960400	33.968150	
5	15.952650	26.680000	24.960400	33.968150	
.	.	.	.	.	
.	.	.	.	.	
45	15.952650	24.300000	24.960400	33.968150	
46	15.952650	29.540000	24.960400	33.968150	
47	15.952650	27.110000	24.960400	33.968150	
48	15.952650	21.500000	24.960400	33.968150	
49	15.952650	23.890000	24.960400	33.968150	
50	15.952650	24.330000	24.960400	33.968150	

58 CAPÍTULO 3. CARTAS DE CONTROL PARA OBSERVACIONES INDIVIDUALES

Observaciones de una Normal(25,9)
sigma estimado por s/c_4

Procedimiento SHEWHART

Resumen del diagrama de medidas individuales para x					
-----Límites 3 Sigma con n=2 para rango móvil-----					
obs	Límite inferior	Rango móvil	de la media	Límite superior	Límite sobrepasado
1	0	.	3.3880523	11.067181	
2	0	4.660000	3.3880523	11.067181	
3	0	3.280000	3.3880523	11.067181	
4	0	3.610000	3.3880523	11.067181	
5	0	3.850000	3.3880523	11.067181	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
46	0	5.240000	3.3880523	11.067181	
47	0	2.430000	3.3880523	11.067181	
48	0	5.610000	3.3880523	11.067181	
49	0	2.390000	3.3880523	11.067181	
50	0	0.440000	3.3880523	11.067181	

Carta para observaciones individuales y para Rangos Moviles

Observaciones de una Normal(25,9)
sigma estimado por s/c_4

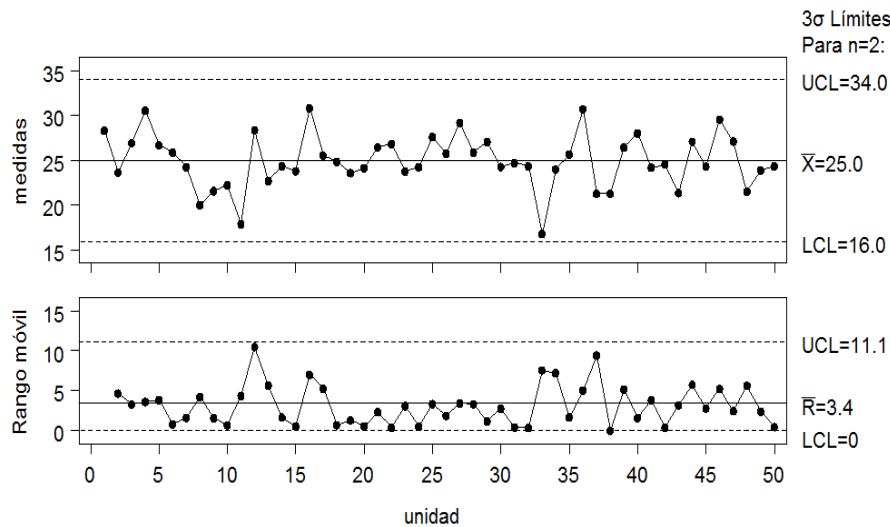


FIGURA 3.5 Carta $\bar{X}$ y de rangos móviles para observaciones individuales. σ estimado como s/c_4 .

De las cartas de control $\bar{X}$ y MR de la figura 3.5, en las cuales se empleó como estimación de sigma a s/c_4 , se puede observar cómo todos los puntos caen dentro de los límites de control (estas cartas contienen todas las observaciones), mientras que estas cartas fueron capaces de detectar algunas observaciones fuera de control usando como estimación

de sigma a $\overline{MR}/d_2$. De este ejemplo se puede ratificar la afirmación hecha al comienzo del ejemplo: el estimador $\overline{MR}/d_2$ se prefiere en la Fase I.

3.2 Cartas de medias móviles

Las cartas de control de medias móviles son otra alternativa para decidir si el proceso está en estado de control, y para detectar desviaciones en la media del proceso.

Para las cartas de control de medias móviles, los límites de control son:

$$\bar{\bar{x}} \pm \frac{3}{\sqrt{n}} \frac{\overline{MR}}{d_2}$$

con

$\bar{\bar{x}}$: promedio de las medias móviles

n : número de observaciones con las cuales se calcula el promedio móvil

$\overline{MR}/d_2$: estimador de σ (d_2 obtenido usando $n = 2$)

El procedimiento para obtener las medias móviles, que son graficadas en las cartas, es como sigue. Si se usan medias de tamaño 5, la primera media será el promedio de las 5 primeras observaciones, la segunda media será el promedio de las observaciones 2 a 6, la tercera será el promedio de las observaciones 3 a 7, y así sucesivamente. Uno de los problemas que tiene esta aproximación radica en que cualquiera de las primeras cuatro observaciones podría indicar que el proceso está fuera de control, pero los primeros cuatro datos no son usados hasta que el quinto esté disponible. Una alternativa sería graficar la primera observación contra el límite de control para observaciones individuales, y luego calcular los límites de control basados en i observaciones, $i = 2, 3, 4$. Obviamente, esto significa que los límites de control sean diferentes para las cinco primeras observaciones graficadas.

Las cartas de control de medias móviles tienen algunas limitaciones de uso cuando se trata de observaciones individuales (ver Ryan, 2000).

3.2.1 Continuación del ejemplo anterior

Con el siguiente programa SAS se obtiene la carta de control de medias móviles, la cual es posible con el argumento MACHART del PROC MACONTROL. Si $n_i = n$, es decir, los tamaños de los subgrupos son iguales, los límites quedan

$$\overline{\overline{X}} - \frac{k\hat{\sigma}}{\sqrt{n * \min(i, w)}}$$

donde w es el ‘span’ para el promedio móvil (número de términos en el promedio móvil) y $\min(i, w)$ es el mínimo entre w e i , que es el número de la observación. Es evidente que los límites de las primeras cuatro observaciones son distintos, como se observa en la figura 3.6.

A continuación se presenta el programa en SAS para la elaboración de la carta de control de medias móviles para los mismos datos de la tabla 3.1.

Código SAS 3.7

```
title 'Carta de Media Movil para datos de tabla 5.1';
title2 'span=5 y sigma estimado como MR/d2';
symbol v=dot c=black;
proc macontrol data=tabla5_1;
machart x*obs / span=5 LIMITN=1 sigma0=2.7911 ndecimal=2
tableall;
run;
```

Note que con la opción $LIMITN = 1$ indica el tamaño de grupo para todos los subgrupos, en este caso $n = 1$, es decir, se trata de observaciones individuales. Además, $\sigma_0 = 2,7911$ indica al SAS que este es el valor de sigma con el cual debe calcular los límites. Este valor resulta de

$$\frac{\overline{MR}}{d_2(2)} = \frac{3,1484}{1,128} = 2,7911 \quad (\text{ver página 63})$$

Si el valor de $\hat{\sigma}$ no es especificado, SAS lo toma por defecto igual a s/c_4 . Así, los resultados son como siguen:

Salida SAS 3.7

Carta de control de Media Móvil para datos de tabla 5.1
span=5 y sigma estimado como MR/d2
Procedimiento MACONTROL

Parámetros de promedio móviles

Sigmas 3
Span 5
Tamaño muestral nominal 1

Resumen del diagrama de media móvil para x

obs	Tamaño de la muestra del subgrupo	Límite inferior	Media Móvil	Media subgrupo
1	1	16.587700	28.300000	28.300000
2	1	19.040183	25.970000	23.640000
3	1	20.126673	26.286667	26.920000
4	1	20.774350	27.347500	30.530000
5	1	21.216346	27.214000	26.680000
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
45	1	21.216346	24.292000	24.300000
46	1	21.216346	25.364000	29.540000
47	1	21.216346	25.878000	27.110000
48	1	21.216346	25.908000	21.500000
49	1	21.216346	25.268000	23.890000
50	1	21.216346	25.274000	24.330000

Resumen del diagrama de media móvil para x

obs	Media Promedio	Límite superior	Límite sobrepasado
1	24.961000	33.334300	
2	24.961000	30.881817	
3	24.961000	29.795327	
4	24.961000	29.147650	
5	24.961000	28.705654	
.	.	.	
.	.	.	
.	.	.	
11	24.961000	28.705654	Inferior
.	.	.	
.	.	.	
.	.	.	
45	24.961000	28.705654	
46	24.961000	28.705654	
47	24.961000	28.705654	
48	24.961000	28.705654	
49	24.961000	28.705654	
50	24.961000	28.705654	

De este ejemplo se puede ver cómo las cartas de control de medias móviles son una alternativa para controlar la media de un proceso, y pueden ser útiles en la Fase I para tratar de llevar el proceso al estado bajo control. Este caso sugiere al ingeniero tratar de identificar los aparentes cambios de media en distintos periodos.

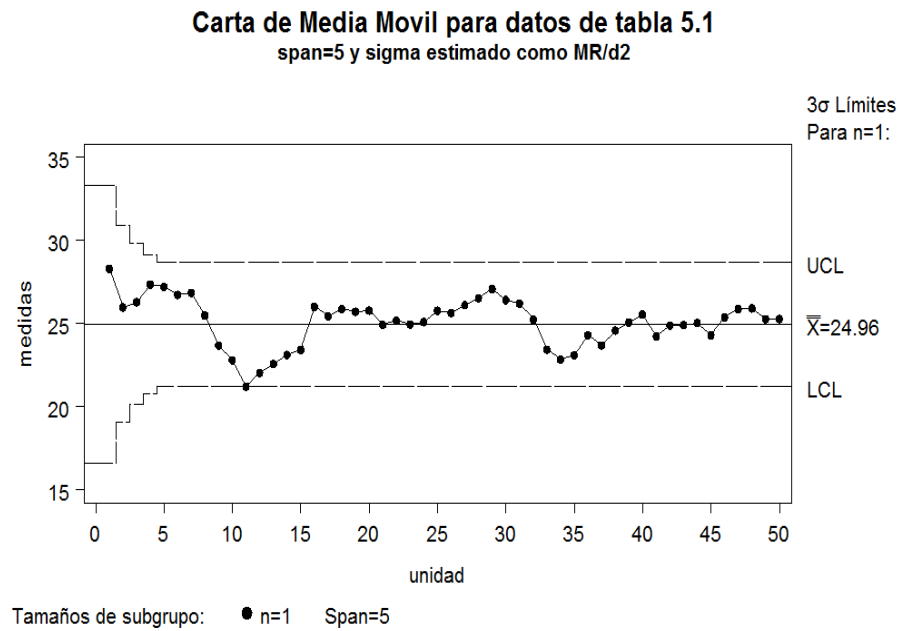


FIGURA 3.6 Carta de control de medias móviles.

CAPÍTULO 4

Cartas de control multivariadas

4.1 ¿Por qué control multivariado?

A nivel univariado, las cartas Shewhart, CUSUM y EWMA han sido ampliamente utilizadas para monitorear los parámetros del proceso que son de interés: medias, varianzas, número de no conformidades. La idea básica en este contexto es la identificación de causas especiales sobre variables que son independientes y que no hayan sido influidas por el comportamiento de otras variables del proceso. Sin embargo, existen muchas situaciones en las que resulta necesario monitorear simultáneamente dos o más características de calidad; por ejemplo, en procesos desarrollados en la industria química y farmacéutica, en los cuales es común encontrar variables de entrada, de proceso y de salida interrelacionadas. En tales casos el control multivariado permite evaluar en forma apropiada el desempeño de un proceso con base en el comportamiento de un conjunto de variables correlacionadas, las cuales no deben ser evaluadas individualmente sino como grupo.

4.1.1 Procesos univariados *versus* procesos multivariados

¿Qué puede suceder si en un proceso multivariado se utilizara un esquema de control univariado para cada una de las variables por controlar?

Suponga por ejemplo que se monitorean por separado las medias de un proceso p -variado de variables independientes distribuidas $N(\mu_i, \sigma_i^2)$, $i = 1, \dots, p$; $p > 1$, estableciendo límites independientes de control en cada caso con una probabilidad de error tipo I de α ; esto resulta equivalente a la realización de p pruebas de hipótesis simultáneas del tipo $H_0 : \mu_{i,t} = \mu_{i,0}$ vs. $H_0 : \mu_{i,t} \neq \mu_{i,0}$. Luego, la verdadera probabilidad de cometer error tipo I en el control conjunto será $P(\text{al menos una falsa alarma}) = 1 - P(\text{ninguna falsa alarma}) = 1 - (1 - \alpha)^p > 1 - \alpha$ y la probabilidad de que todas las p medias se encuentren en control será de $(1 - \alpha)^p$. Ahora se ilustrará numéricamente: suponga $\alpha = 0,0027$ y $p = 200$, entonces $P(\text{error tipo I}) = 0,42$; esto quiere decir que hay una tasa de falsa alarma del 42 %, lo cual es algo intolerable en un proceso de producción y termina siendo algo parecido a la metáfora del ‘pastorcito mentiroso’: ¡Nadie le cree! Ahora bien, si las variables son correlacionadas, como ocurre con frecuencia en la práctica, la tasa de falsas alarmas crece.

Cuando las p características de calidad no son independientes y se distribuyen conjuntamente $N_p(\mu, \Sigma)$, donde μ es el vector de medias y Σ es la matriz de varianzas covarianzas, no son válidas las dos expresiones dadas para las probabilidades de error tipo I conjunto y del estado de medias en control, respectivamente, y, en general, no existe una manera fácil de medir la distorsión en el procedimiento de control conjunto; es decir, una alta correlación entre las variables puede hacer imposible establecer la tasa total de falsas alarmas producidas por un gran número de cartas de control univariadas. De nuevo cabe señalar que se puede concluir erróneamente sobre el estado del proceso multivariado cuando se recurre a cartas de control univariadas para monitorear simultáneamente un conjunto de variables, entre las cuales dos o más pueden no ser independientes estadísticamente.

Para entender mejor lo anterior, considere el caso sencillo de un proceso normal bivariado, donde X_1 y X_2 son las variables de interés, suponiendo una correlación positiva entre estas variables; la región de control

multivariada tendrá forma elíptica (luego se verá por qué) orientada según el signo de la correlación, en tanto que la región de control definida por el uso de dos esquemas univariados aislados, al desconocer la asociación lineal entre las dos variables, tendrá la apariencia de un rectángulo, lo cual conduce a conclusiones erróneas respecto al estado de control del proceso (ver figura 4.1).

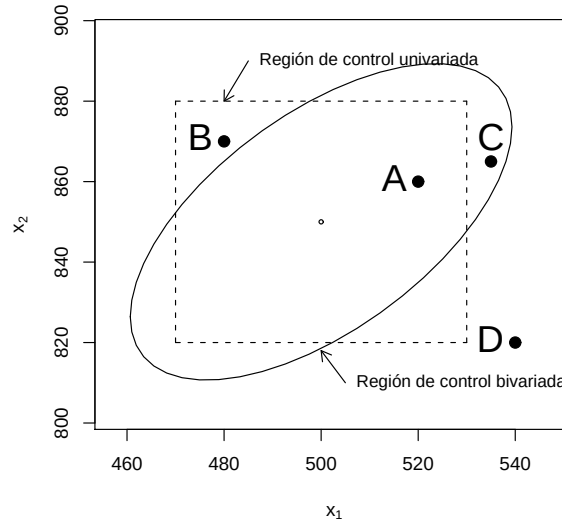


FIGURA 4.1 Región de control bivariada *vs.* región de control con dos cartas univariadas. Punto A: Cae en la región de control de ambos procedimientos. Punto B: El esquema univariado indica que el proceso está en control cuando de hecho no lo está. Punto C: El esquema univariado indica que el proceso está fuera de control cuando de hecho está en control. Punto D: Ambos esquemas indican que el proceso está fuera de control.

En conclusión, al existir una estructura de correlación entre las variables, estas deben ser evaluadas como un todo. La correlación entre variables de un proceso puede ser debida a (Mason & Young, 2002):

- Asociación: Por efecto de variables no observables. En este caso, no existe una relación de causa-efecto entre dos variables, sino que sus valores presentan una covariación debido a que son afectadas por una tercera, desconocida o imposible de medir directamente.
- Causalidad: Por relación físico-química de causa-efecto.

En cualquier caso, es necesario tener en cuenta las asociaciones que existen entre las variables del proceso, con una aproximación multivariada que dé cuenta de la dependencia y controle efectivamente la tasa de falsas alarmas. Esto no significa que se descarta el uso de cartas de control univariadas, sino más bien que estas se utilizan como un complemento en la identificación de señales en el proceso multivariado.

4.1.2 Características deseables de un procedimiento de control multivariado

Los sistemas avanzados de control de procesos sobre una gran variedad de variables son usados básicamente para asegurar que dichas variables se mantengan cerca de un valor objetivo o dentro de un rango operacional aceptable; en ese sentido, reducen la variabilidad total del proceso (si no se llega a sobrecontrolar); sin embargo, de ninguna manera garantizan que el proceso satisfaga un conjunto de condiciones de línea base y tampoco permiten identificar las causas de las perturbaciones ocurridas en el proceso. Este tipo de control debe ser complementado con el SPC para detectar los cambios en los modelos de control automático establecidos, así como la detección de causas. Por tanto no se debe pensar que el ajuste automático es sinónimo de control multivariado.

Mason & Young (2002) señalan las siguientes características deseables para un esquema de control multivariado:

1. El monitoreo estadístico a través de cartas de control debe ser de fácil construcción y útil al identificar tendencias en el proceso.
2. Cuando ocurre un punto fuera de control, debe ser fácil determinar la causa en términos del conjunto de variables que determinan dicha señal.
3. El procedimiento debe ser flexible en su aplicación.
4. El procedimiento debe ser sensible a cambios pequeños pero persistentes en el proceso. En algunos procesos multivariados, cambios pequeños pueden conducir a grandes riesgos y pérdidas.
5. El procedimiento debería ser capaz de monitorear el proceso tanto en línea como fuera de línea.

Un procedimiento de control que cumple con las anteriores características es la carta de control T^2 de Hotelling (ver figura 4.2), basado en el estadístico T^2 . En esta carta la interpretación de señales requiere un procedimiento para aislar la contribución de cada variable o grupo de variables; una señal puede ser atribuible a una sola de las variables o a una relación entre dos o más variables.

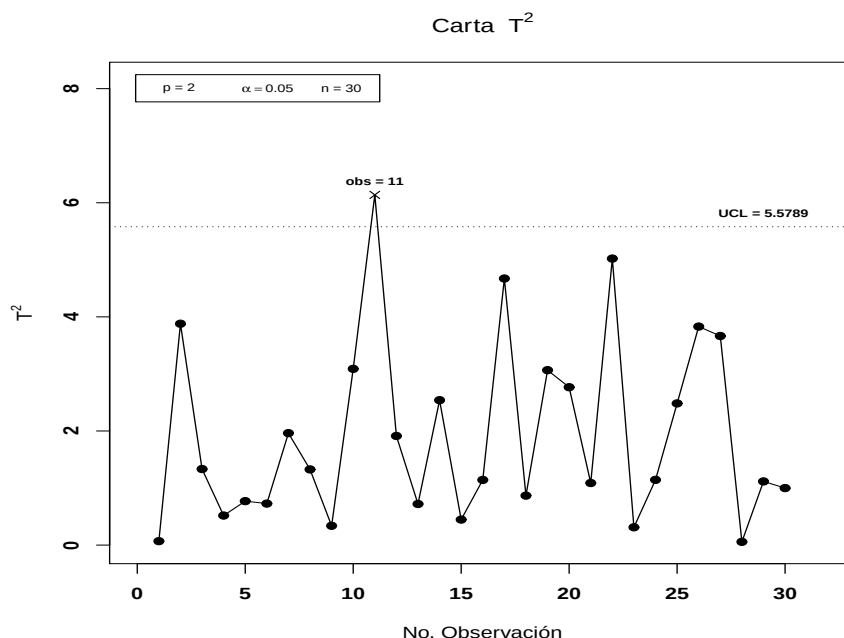


FIGURA 4.2 Carta de control T^2 en la Fase I, con 30 observaciones históricas de un proceso bivariado.

Se verá un procedimiento para interpretar señales capaz de descomponer el estadístico T^2 en componentes independientes. Estas componentes pueden clasificarse en dos tipos: un tipo de componentes determinará la contribución de las variables individuales al valor de la observación T^2 que generó una señal, mientras que el otro tipo de componentes chequearán las relaciones entre dos o más variables.

Otras cartas propuestas para el control de calidad multivariado son: las cartas CUSUM multivariadas o MCUSUM (Woodall y Ncube, 1985; Crosier, 1986, 1988; Healy, 1987), las cartas EWMA multivariadas o

MEWMA (Lowry *et al.*, 1992), y una propuesta de cartas de perfiles denominadas MP Charts (Fuchs y Benjamin, 1994). Para una revisión sobre cartas de control multivariadas, ver también a Lowry y Montgomery (1995), Kourty y MacGregor (1996) y Mason y Young (2002). También se han desarrollado propuestas para el caso de control con observaciones multivariadas correlacionadas en el tiempo (ver, por ejemplo, Mason *et al.*, 2003) y para el control de procesos que operan por lotes (ver, por ejemplo, Mason *et al.*, 2001).

4.2 Aspectos preliminares: estudio del T^2 de Hotelling

4.2.1 Variables y observaciones

El análisis multivariado trata con datos que contienen observaciones sobre dos o más variables medidas sobre un conjunto de objetos o individuos. En este sentido, las observaciones están constituidas por vectores de dimensión p , donde p corresponde al número de variables simultáneamente observadas sobre cada individuo u objeto.

$$X_{n \times p} = \begin{matrix} & \text{var}_1 & \text{var}_2 & \cdots & \text{var}_p \\ \begin{matrix} \text{obs}_1 \\ \text{obs}_2 \\ \vdots \\ \text{obs}_n \end{matrix} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \end{matrix}$$

Por ejemplo, considere la tabla 4.1 de datos artificiales, en la cual se presenta información sobre las curvas de absorción para diez diferentes muestras medidas en siete longitudes de onda.

En esta tabla hay diez observaciones, cada una de las cuales proporciona información sobre los valores de siete variables, es decir, las siete longitudes de onda.

TABLA 4.1 Datos de absorción.

No. Longitud de onda							
Obs.	1	2	3	4	5	6	7
1	0,5	1,0	1,5	1,0	0,5	1,5	2,5
2	1,0	2,0	3,0	2,0	1,0	3,0	5,0
3	1,5	3,0	4,5	3,0	1,5	4,5	7,5
4	2,0	4,0	6,0	4,0	2,0	6,0	10,0
5	2,5	5,0	7,5	5,0	2,5	7,5	12,5
6	3,0	6,0	9,0	6,0	3,0	9,0	15,0
7	3,5	7,0	10,5	7,0	3,5	10,5	17,5
8	4,0	8,0	12,0	8,0	4,0	12,0	20,0
9	4,5	9,0	13,5	9,0	4,5	13,5	22,5
10	5,0	10,0	15,0	10,0	5,0	15,0	25,0

Fuente: Jackson (1991), p. 75.

4.2.2 Matriz de datos y estadísticos muestrales multivariados

Un conjunto de datos que comprende n observaciones tomadas sobre p variables puede ser representado en una matriz $\mathbf{X}$ de orden $n \times p$, que se denominará matriz de datos. Así, las columnas representan las variables, y las filas los individuos o vectores de observación.

Considere una muestra aleatoria de n observaciones p -variadas, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$; sea $\mathbf{x}_i^t$ la transpuesta¹ de la i -ésima fila (observación i) de la matriz $\mathbf{X}$, provenientes de una distribución multivariada que tiene como media al vector

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad (4.1)$$

¹t denota la transpuesta de un vector o matriz. Por convención, los vectores se consideran siempre como vectores columna, de modo que $\mathbf{x}_i^t$ denota el vector fila.

y matriz de varianzas covarianzas

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_p^2 \end{pmatrix} \quad (4.2)$$

donde σ_{ij} corresponde a la covarianza entre las variables i y j y σ_i^2 es la varianza de la variable i . Obsérvese que la matriz $\mathbf{\Sigma}$ es simétrica. El i -ésimo vector de observación corresponde a:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad (4.3)$$

Si las n filas $\mathbf{x}_i$ conforman una muestra aleatoria de vectores de $\mathbb{R}^p$, de una distribución p -variada, el vector de medias muestrales correspondiente es:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}, \text{ donde } \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (4.4)$$

representa la media aritmética de la j -ésima variable. También se puede escribir la media muestral en términos de la matriz de datos:

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X}^t \mathbf{1}_{n \times 1} \quad (4.5)$$

donde $\mathbf{1}_{n \times 1}$ es un vector columna de dimensión n con todos sus elementos iguales a 1.

La matriz de varianzas covarianzas muestral $\mathbf{S}_n$ (estimador sesgado de $\mathbf{\Sigma}$) está dada por:

$$\begin{aligned} \mathbf{S}_n &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^t \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^t - \bar{\mathbf{x}} \bar{\mathbf{x}}^t \end{aligned} \quad (4.6)$$

o en términos de la matriz de datos:

$$\begin{aligned}
 \mathbf{S}_n &= \frac{1}{n} \mathbf{X}^t \mathbf{X} - \overline{\mathbf{x}} \overline{\mathbf{x}}^t \\
 &= \frac{1}{n} \left[\mathbf{X}^t \mathbf{X} - \frac{1}{n} (\mathbf{X}^t \mathbf{1}_{n \times 1}) (\mathbf{1}_{1 \times n}^t \mathbf{X}) \right] \\
 &= \frac{1}{n} \mathbf{X}^t \left[\mathbf{I}_{n \times n} - \frac{1}{n} \mathbf{1}_{n \times 1} \mathbf{1}_{1 \times n}^t \right] \mathbf{X} \\
 &= \frac{1}{n} \mathbf{X}^t \mathbf{H} \mathbf{X}
 \end{aligned} \tag{4.7}$$

donde $\mathbf{H}$ de orden $n \times n$ es conocida como la matriz de centramiento, y es tal que:

1. $\mathbf{H}^t = \mathbf{H}$, es decir, es simétrica
2. $\mathbf{H}^2 = \mathbf{H}$, es decir, es idempotente, y
3. $\mathbf{H}$ es semidefinida positiva

Sea $\mathbf{a} \in \mathbb{R}^p$, entonces:

$$\begin{aligned}
 \mathbf{a}^t \mathbf{S}_n \mathbf{a} &= \frac{1}{n} \mathbf{a}^t \mathbf{X}^t \mathbf{H} \mathbf{X} \mathbf{a} \\
 &= \frac{1}{n} \mathbf{a}^t \mathbf{X}^t \mathbf{H}^t \mathbf{H} \mathbf{X} \mathbf{a}
 \end{aligned} \tag{4.8}$$

Sea $\mathbf{Y} = \mathbf{H} \mathbf{X} \mathbf{a}$, $\mathbf{Y} \in \mathbb{R}^n$, entonces

$$\begin{aligned}
 \mathbf{a}^t \mathbf{S}_n \mathbf{a} &= \frac{1}{n} \mathbf{Y}^t \mathbf{Y} \\
 &= \frac{1}{n} \|\mathbf{Y}\|^2 \geq 0
 \end{aligned} \tag{4.9}$$

por tanto, $\mathbf{S}_n$ es semidefinida positiva, donde $\|\cdot\|$ representa la norma o longitud de un vector en $\mathbb{R}^n$. Para datos continuos, generalmente $\mathbf{S}_n$ es definida positiva si $n \geq p + 1$.

Ahora considere a:

$$\begin{aligned}
 \mathbf{S}_u &= \frac{1}{n-1} \mathbf{X}^t \mathbf{H} \mathbf{X} \\
 &= \frac{n}{n-1} \mathbf{S}_n
 \end{aligned} \tag{4.10}$$

este es un estimador insesgado de la matriz de varianzas covarianzas.

También se puede expresar a $\mathbf{S}_n$ en términos de su ij -ésimo elemento:

$$S_{ij} = \frac{1}{n} \sum_{r=1}^n (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) \quad (4.11)$$

que corresponde al estimador sesgado de σ_{ij} . Luego, para $i = j$ se tiene que:

$$S_{ii} = S_i^2 = \frac{1}{n} \sum_{r=1}^n (x_{ri} - \bar{x}_i)^2 \quad (4.12)$$

este último es un estimador sesgado de σ_i^2 . La forma de la matriz $\mathbf{S}_n$, que es simétrica, es la siguiente:

$$\mathbf{S}_n = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{12} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p} & S_{2p} & \cdots & S_{pp} \end{pmatrix} \quad (4.13)$$

Para $\mathbf{S}_u$ las componentes están dadas por:

$$S_{u,ij} = \frac{1}{n-1} \sum_{r=1}^n (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) \quad (4.14)$$

Considere ahora

$$r_{ij} = \frac{S_{ij}}{S_i S_j} \quad (4.15)$$

que corresponde a un estimador del coeficiente de correlación entre las variables i y j . Sea $\mathbf{R} = \{r_{ij}\}$ la matriz de correlaciones muestrales; se puede demostrar que esta es semidefinida positiva.

4.2.3 Medidas de dispersión multivariadas

Existen dos estadísticos que permiten obtener medidas multivariadas de dispersión asignando un solo valor numérico a la variación expresada por $\mathbf{S}_n$:

- La varianza generalizada, que corresponde a $|\mathbf{S}_n|$, donde $|\cdot|$ es la función determinante.

- La variación total, dada por $tr(\mathbf{S}_n)$, donde $tr(\cdot)$ denota la función traza.

En ambas medidas, valores grandes dan indicio de una alta dispersión de $X_1, X_2, \dots, X_p$ alrededor de $\bar{\mathbf{X}}$. Ahora bien, en el caso de la varianza generalizada, un valor cercano a 0 indica que al menos una de las columnas de la matriz de desviaciones ($X_{n \times p} - \mathbf{1}_{n \times 1} \bar{x}_{1 \times p}$) puede ser expresada como una combinación lineal de las otras columnas. En el caso de la variación total, que es la suma del total de las varianzas muestrales, un valor pequeño indica que las varianzas muestrales son pequeñas. Sin embargo, cada medida refleja aspectos diferentes de la variabilidad de los datos. La varianza generalizada desempeña un papel importante en la estimación de máxima verosimilitud, en tanto que la variación total es un concepto útil en el análisis de componentes principales.

4.2.4 Combinaciones lineales

Las combinaciones lineales pueden simplificar la estructura de la matriz de varianzas covarianzas haciendo más directa la interpretación de los datos. Sea una combinación lineal

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \dots + a_p x_{ip} \quad i = 1, 2, \dots, n \quad (4.16)$$

donde los a_i son constantes reales. Note que y_i es un escalar. Se puede hallar la media de las n combinaciones lineales, así:

$$\bar{y} = \mathbf{a}_{1 \times p}^t \bar{\mathbf{x}}_{p \times 1}, \text{ donde } \mathbf{a}^t = (a_1, \dots, a_p) \quad (4.17)$$

o bien, definiendo el vector que contiene las n combinaciones lineales

$$\mathbf{Y} = \mathbf{X}_{n \times p} \mathbf{a}_{p \times 1} \quad (4.18)$$

entonces

$$\bar{y} = \frac{1}{n} \mathbf{a}^t \mathbf{X}^t \mathbf{1}_{n \times 1} \quad (4.19)$$

y la varianza de las n combinaciones lineales corresponde a

$$\begin{aligned} S_y^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{a}^t (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^t \mathbf{a} \\ &= \mathbf{a}^t \mathbf{S}_n \mathbf{a} \end{aligned} \quad (4.20)$$

También se puede hablar de transformaciones lineales q dimensionales; en ese caso:

$$\mathbf{y}_{i_{q \times 1}} = \mathbf{A}_{q \times p} \mathbf{x}_i + \mathbf{b}_{q \times 1} \quad i = 1, 2, \dots, n \quad (4.21)$$

donde $\mathbf{A}$ y $\mathbf{b}$ son matrices de constantes reales. Por tanto, las n transformaciones lineales q dimensionales se pueden escribir matricialmente como:

$$\mathbf{Y}_{n \times q} = \mathbf{X}_{n \times p} \mathbf{A}_{p \times q}^t + \mathbf{1}_{n \times 1} \mathbf{b}_{1 \times q}^t \quad (4.22)$$

Usualmente $q \leq p$. El vector de medias de $\mathbf{Y}_{n \times q}$ está dado por:

$$\begin{aligned} \bar{\mathbf{Y}}_{q \times 1} &= \frac{1}{n} \mathbf{Y}_{q \times n}^t \mathbf{1}_{n \times 1} \\ &= \frac{1}{n} (\mathbf{X}_{n \times p} \mathbf{A}_{p \times q}^t + \mathbf{1}_{n \times 1} \mathbf{b}_{1 \times q}^t)^t \mathbf{1}_{n \times 1} \\ &= \frac{1}{n} (\mathbf{A}_{q \times p} \mathbf{X}_{p \times n}^t \mathbf{1}_{n \times 1} + \mathbf{b}_{q \times 1} \mathbf{1}_{1 \times n}^t \mathbf{1}_{n \times 1})^t \\ &= \mathbf{A} \bar{\mathbf{X}} + \mathbf{b} \end{aligned} \quad (4.23)$$

La matriz de varianzas covarianzas de $\mathbf{Y}_{n \times q}$ es (tener en cuenta que $\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_{n \times 1} \mathbf{1}_{1 \times n}^t$, $\mathbf{1}_{1 \times n}^t \mathbf{H}_{n \times n} = \mathbf{0}_{1 \times n}$ y $\mathbf{H}_{n \times n} \mathbf{1}_{n \times 1} = \mathbf{0}_{n \times 1}$):

$$\begin{aligned} \mathbf{S}_Y &= \frac{1}{n} \mathbf{Y}^t \mathbf{H} \mathbf{Y} \\ &= \frac{1}{n} (\mathbf{X} \mathbf{A}^t + \mathbf{1} \mathbf{b}^t)^t \mathbf{H} (\mathbf{X} \mathbf{A}^t + \mathbf{1} \mathbf{b}^t) \\ &= \frac{1}{n} \mathbf{A} \mathbf{X}^t \mathbf{H} \mathbf{X} \mathbf{A}^t \\ &= \mathbf{A} \left(\frac{1}{n} \mathbf{X}^t \mathbf{H} \mathbf{X} \right) \mathbf{A}^t \\ &= \mathbf{A} \mathbf{S}_n \mathbf{A}^t \end{aligned} \quad (4.24)$$

4.2.5 Transformaciones

Para las siguientes secciones, considere una muestra aleatoria de n observaciones p -variadas, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$.

1. **Transformación de estandarización.** Sea $\mathbf{Y}_i$ la variable dada por:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{D}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad i = 1, 2, \dots, n \\ \mathbf{D} &= \text{diag}(S_i) \\ S_i &= \sqrt{\frac{1}{n} \sum_{r=1}^n (x_{ri} - \bar{x}_i)^2} \end{aligned} \quad (4.25)$$

Esta transformación escala las variables centradas de modo que tengan varianza unitaria. Fácilmente, se puede mostrar que

$$\bar{\mathbf{Y}}_{n \times 1} = \mathbf{0}_{n \times 1} \quad (4.26)$$

teniendo como referente el caso univariado donde $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

A esta operación se le llama centramiento de los datos en el origen.

También se puede mostrar que la matriz de varianzas covarianzas del vector $\mathbf{Y}_i$

$$\mathbf{S}_{\mathbf{Y}} = \mathbf{D}^{-1} \mathbf{S}_n \mathbf{D}^{-1} = \mathbf{R} \quad (4.27)$$

Recuerde que $\mathbf{R}$ es la matriz de correlación, y por consiguiente, tiene unos en la diagonal principal, lográndose así el propósito de varianza unitaria.

2. **Transformación de componentes principales.** Por el teorema de descomposición espectral del álgebra lineal $\mathbf{S}_n$, se puede escribir:

$$\mathbf{S}_n = \mathbf{G} \mathbf{\Lambda} \mathbf{G}^t \quad (4.28)$$

y además

$$\mathbf{\Lambda} = \mathbf{G}^t \mathbf{S}_n \mathbf{G} \quad (4.29)$$

donde $\mathbf{G}_{p \times p}$ es una matriz ortogonal formada a partir de los vectores propios de $\mathbf{S}_n$, $\mathbf{\Lambda}$ es una matriz diagonal de valores propios de $\mathbf{S}_n$, con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

Sea $\mathbf{W}_i$ el vector de transformación de la observación i dado por:

$$\mathbf{W}_i = \mathbf{G}^t (\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, 2, \dots, n \quad (4.30)$$

Claramente, $\bar{\mathbf{W}}_{n \times 1} = \mathbf{0}$, y la matriz de varianzas covarianzas del vector $\mathbf{W}_i$ es

$$\mathbf{S}_{\mathbf{W}} = \mathbf{G}^t \mathbf{S}_n \mathbf{G} = \mathbf{\Lambda} \quad (4.31)$$

Las p columnas de la matriz $\mathbf{W}_{n \times p}$ son llamadas componentes principales, y representan combinaciones lineales incorrelacionadas de las p variables originales; es decir, la varianza de las componentes principales y que según (4.31) tienen, respectivamente, varianza igual a $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.

En este caso, la varianza generalizada $|\mathbf{S}_n| = |\mathbf{\Lambda}| = \prod_{i=1}^p \lambda_i$ y la variación total $tr(\mathbf{S}_n) = tr(\mathbf{\Lambda}) = \sum_{i=1}^p \lambda_i$

3. **Transformación de Mahalanobis.** Si $\mathbf{S}_n > 0$ (definida positiva) tiene una única raíz cuadrada definida positiva $\mathbf{S}_n^{1/2}$. Sea

$$\mathbf{Z}_i = \mathbf{S}_n^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, 2, \dots, n \quad (4.32)$$

Es claro que $\bar{\mathbf{Z}}_{n \times 1} = \mathbf{0}$, y la matriz de varianzas covarianzas del vector $\mathbf{Z}_i$ es

$$\mathbf{S}_{\mathbf{Z}} = \mathbf{S}_n^{-1/2} \mathbf{S}_n \mathbf{S}_n^{-1/2} = \mathbf{I}_p \quad (4.33)$$

de modo que esta transformación centra y esfera a los datos (es decir, elimina la correlación entre las variables).

Cabe anotar que las anteriores tres transformaciones también se definen usando $\mathbf{S}_u$ en lugar de $\mathbf{S}_n$.

4.2.6 Visión geométrica de la matriz de datos

Considere la muestra aleatoria de tamaño n de observaciones p -variadas, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$; sea $\mathbf{x}_{n \times p}$ la correspondiente matriz de datos dada por:

$$\mathbf{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad (4.34)$$

Espacio objeto

Las columnas de $\mathbf{X}$ pueden ser vistas como p puntos en un espacio n -dimensional llamado *espacio R* o *espacio objeto*.

Sea la matriz centrada $\mathbf{Y} = \mathbf{H}\mathbf{X}$ donde las nuevas p variables tienen media cero; entonces r_{ij} la correlación entre las columnas i y j de $\mathbf{X}$ se puede ver como:

$$\cos \theta_{ij} = \frac{\mathbf{Y}_{(i)}^t \mathbf{Y}_{(j)}}{\|\mathbf{Y}_{(i)}\| \|\mathbf{Y}_{(j)}\|} = \frac{S_{ij}}{S_i S_j} = r_{ij} \quad (4.35)$$

es decir, la correlación muestral entre dos variables corresponde al coseno del ángulo de los vectores variables centrados.

Espacio variable

Las n filas de $\mathbf{X}$ se pueden considerar como n puntos en un espacio p dimensional llamado *espacio Q* o *espacio variable*.

Una forma de comparar dos filas (observaciones) $\mathbf{x}_i$ y $\mathbf{x}_j$ es a través de la distancia euclidiana de estos vectores:

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^t (\mathbf{x}_i - \mathbf{x}_j) \quad (4.36)$$

Otra forma de comparación es transformar los datos y mirar la distancia euclidiana entre las filas transformadas, por ejemplo, usando la transformación de Mahalanobis, que origina la llamada distancia de Mahalanobis:

$$D_{ij}^2 = \|\mathbf{Z}_i - \mathbf{Z}_j\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^t \mathbf{S}_n^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (4.37)$$

4.3 La distribución normal multivariada

En la práctica existen muchos problemas multivariados en los cuales la teoría basada en la distribución normal multivariada funciona apropiadamente. En el contexto de control de calidad, cuando se tiene un conjunto de variables interrelacionadas para la construcción de cartas de control, se utiliza el supuesto de multinormalidad de las variables del proceso. Se presenta a continuación una breve descripción de la distribución normal multivariada y algunas de sus propiedades.

4.3.1 Densidad normal multivariada

Para el caso univariado ($p = 1$), la función de densidad normal es:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right], \quad -\infty < x < \infty$$

El término

$$\frac{(x - \mu)^2}{\sigma^2} = (x - \mu) (\sigma^2)^{-1} (x - \mu)$$

mide el cuadrado de la distancia de x a μ en unidades estándar.

Por analogía, se mostrará la forma de la función de densidad de la normal multivariada. Para el caso p -variado donde $\mathbf{x}$ es un vector de observaciones $p \times 1$, esta expresión corresponderá a:

$$(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (4.38)$$

con $\boldsymbol{\mu}$ el vector de medias de dimensión $p \times 1$ (ver ecuación (4.1)) y $\boldsymbol{\Sigma}$ la matriz de varianzas covarianzas de dimensión $p \times p$ (ver ecuación (4.2)).

La constante de normalización $(2\pi)^{-1/2} (\sigma^2)^{-1/2}$ debe ser cambiada por aquella con la cual el volumen bajo la superficie de la función de densidad multivariada es igual a 1. Esta constante es $(2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2}$, donde $|\boldsymbol{\Sigma}|$ representa la varianza poblacional generalizada; por tanto, para el vector aleatorio $\mathbf{x} = [x_1, x_2, \dots, x_p]^t \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, la función de densidad normal p -variada es:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \quad (4.39)$$

4.3.2 Distribución normal bivariada

Un caso particular de (4.39) es para $p = 2$, con $\rho_{1,2} = \text{corr}(x_1, x_2)$

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp \left\{ -\frac{1}{2(1-\rho_{12}^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\} \quad -\infty < x_1, x_2 < \infty \quad (4.40)$$

En este caso

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

de donde

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}$$

con $\sigma_{12} = \sigma_1\sigma_2\rho_{12}$; por tanto, la ecuación (4.40) se puede escribir $(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ que es de la forma de la ecuación (4.39).

La figura 4.3 ilustra varias funciones normales bivariadas. Por la forma que se observa, generalmente se le conoce como la “campana de Gauss”.

4.3.3 Contornos de densidad constante

Considere la figura 4.3. Si se corta la superficie con un plano perpendicular al eje $z = f(x_1, x_2)$, el corte resultante sobre dicho plano sería un contorno elíptico, y es tal que comprende todos los pares (x_1, x_2) tales que $(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ es constante. En general, se definen los contornos de densidad constante como la superficie de un elipsoide centrado en $\boldsymbol{\mu}$ donde

$$\left\{ \mathbf{x} \mid \mathbf{x} \in (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2 \right\} \quad (4.41)$$

Los ejes de cada elipsoide de densidad constante están en la dirección de los vectores propios de Σ , y sus longitudes son proporcionales a las

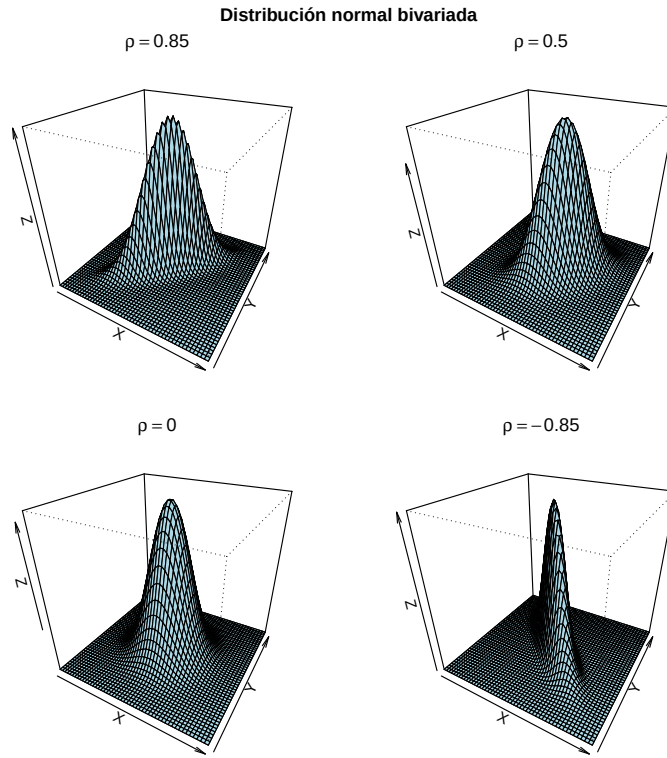


FIGURA 4.3 Distribuciones normales bivariadas, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ y $\sigma_{12} = \rho$.

raíces cuadradas de los valores propios de Σ . Es decir, los ejes de estos elipsoides son $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, donde $\Sigma\mathbf{e}_i = \lambda_i\mathbf{e}_i$, λ_i son los valores propios y $\mathbf{e}_i$ son los vectores propios, con $i = 1, 2, \dots, p$.

Se puede probar que $(\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ converge asintóticamente a una χ_p^2 . Por consiguiente, si se toma $c^2 = \chi_p^2(\alpha)$, donde $\chi_p^2(\alpha)$ es el percentil superior $100\alpha\%$ de la distribución chi cuadrado con p grados de libertad, entonces los contornos obtenidos conducen a contornos que contienen $100(1 - \alpha)\%$ de la probabilidad bajo la densidad normal p -variada. Por analogía, note que en el caso univariado se habla de intervalos de confianza, mientras que en el caso multivariado se hablará de contornos o regiones de confianza.

A continuación se da el programa R para construir los contornos de las elipses de densidad constante. La función para construir dichas elipses

es `confelli`, cuyos parámetros son:

- `mu`: vector de medias
- `Sigma`: matriz de varianzas covarianzas
- `df`: grados de libertad ($n - 2$)
- `level`: nivel de confianza. Por defecto es 0,95

Para emplear la función, se escribe de la siguiente forma: `confelli(mu,Sigma,df,level)`

Código R 4.1

```
confelli<-function(mu,Sigma,df,level=0.95,xlab="",ylab="",
add=T,prec=51,etiqueta=T)
{
  d <- sqrt(diag(Sigma))
  dfvec <- c(2, df)
  phase <- acos(Sigma[1, 2]/(d[1] * d[2]))
  angles <- seq( - (pi), pi, len = prec)
  mult <- sqrt(dfvec[1] * qf(level, dfvec[1], dfvec[2]))
  xpts <- mu[1] + d[1] * mult * cos(angles)
  ypts <- mu[2] + d[2] * mult * cos(angles + phase)
  if(add)lines(xpts, ypts)
  else plot(xpts, ypts, type = "l", lwd=3,xlab = xlab,
ylab = ylab)
  a<-round(runif(1,1,51))
  if(etiqueta==TRUE){
    text(xpts[a], ypts[a],paste(level),adj=c(0.5,0.5),font=2,
cex=0.7)}
}

#Simulando 50 observaciones de una normal bivariada
library(MASS)
mu<-c(1,2)
Sigma<-matrix(c(4,0.95,0.95,1),ncol=2)

datos<-mvrnorm(50,mu,Sigma)
x<-datos[,1]
y<-datos[,2]

#Graficando los 50 pares de puntos (x,y) (ejes sin marcas
de ticks)
plot(x,y,pch=19,xlim=c(min(x)-2,max(x)+2),ylim=c(min(y)-2,
max(y)+2),cex=0.5,
xlab=expression(x[1]),ylab=expression(x[2]),xaxt="n",yaxt="n")
```

```
axis(1,at=1,labels=expression(mu[1]))
axis(2,at=2,labels=expression(mu[2]))

#Graficando contornos de probabilidad normales bivariados
con función confelli
confelli(mu,Sigma,df=48,level=0.30)
confelli(mu,Sigma,df=48,level=0.50)
confelli(mu,Sigma,df=48,level=0.70)
confelli(mu,Sigma,df=48,level=0.90)

#Agregando ejes horizontal y vertical pasando por (mu1,mu2)
abline(h=mu[2],lty=2)
abline(v=mu[1],lty=2)
```

Salida R 4.1

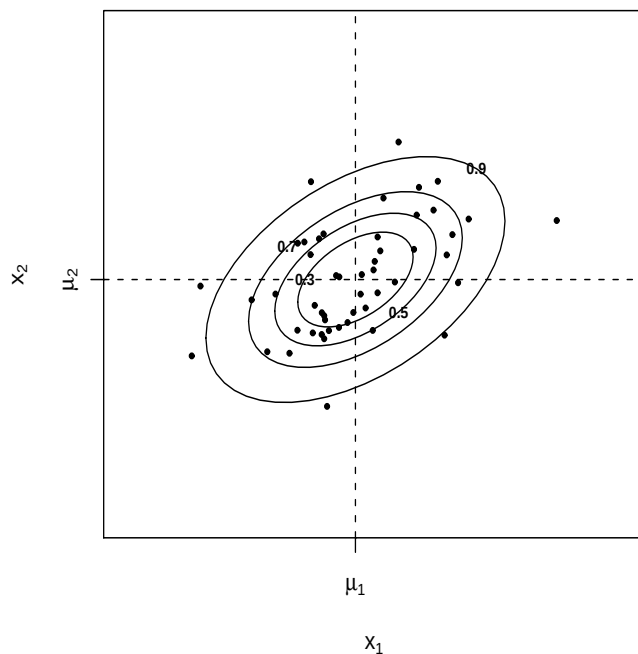


FIGURA 4.4 Contornos de probabilidad del 30 %, 50 %, 70 % y 90 %, de una normal bivariada con $\mu = (1, 2)^t$, $\sigma_1^2 = 4$, $\sigma_2^2 = 1$ y $\rho_{12} = 0,95$.

4.3.4 Algunas propiedades de la distribución normal multivariada

Sea el vector aleatorio $\mathbf{x} = [x_1, x_2, \dots, x_p]^t \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, es decir, se distribuye normal p -variada, con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas covarianzas $\boldsymbol{\Sigma}$. El vector $\mathbf{x}$ tiene las siguientes propiedades

1. Combinaciones lineales de las componentes de $\mathbf{x}$ se distribuyen en forma normal. Se tienen dos casos:

- Considere el vector $\mathbf{a} = [a_1, a_2, \dots, a_p]^t$. La combinación lineal dada por

$$\begin{aligned} \mathbf{a}^t \mathbf{x} &= a_1 x_1 + a_2 x_2 + \dots + a_p x_p, \forall a_i \in \mathbb{R} \\ \mathbf{a}^t \mathbf{x} &\sim N(\mathbf{a}^t \boldsymbol{\mu}, \mathbf{a}^t \boldsymbol{\Sigma} \mathbf{a}), \forall \mathbf{a} \in \mathbb{R}^p \end{aligned} \quad (4.42)$$

da origen a una variable aleatoria normal univariada.

- Considere la matriz

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \dots & \vdots \\ a_{q1} & a_{q2} & \dots & a_{qp} \end{pmatrix}$$

El vector de combinaciones lineales dado por

$$\mathbf{A} \mathbf{x} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p \\ \vdots \\ a_{q1}x_1 + a_{q2}x_2 + \dots + a_{qp}x_p \end{pmatrix} \quad (4.43)$$

$$\mathbf{A} \mathbf{x} \sim N_q(\mathbf{A} \boldsymbol{\mu}, \mathbf{A} \boldsymbol{\Sigma} \mathbf{A}^t), \forall \mathbf{A} \in \mathbb{R}^{q \times p},$$

origina un vector aleatorio que se distribuye en forma normal q -variada. En general, transformaciones lineales del tipo $\mathbf{A} \mathbf{x} + \mathbf{b}$ son normales multivariadas, si $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2. Todos los subconjuntos de componentes de $\mathbf{x}$ tienen distribución normal multivariada. Si se particiona $\mathbf{x}$, su vector de medias $\boldsymbol{\mu}$ y su matriz de covarianza $\boldsymbol{\Sigma}$, se pueden escribir de la siguiente forma

$$\mathbf{x}_{p \times 1} = \begin{pmatrix} \mathbf{x}_{1 \times 1} \\ \mathbf{x}_{2 \times (p-1)} \end{pmatrix}$$

$$\boldsymbol{\mu}_{p \times 1} = \begin{pmatrix} \boldsymbol{\mu}_{1q \times 1} \\ \boldsymbol{\mu}_{2(p-q) \times 1} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{p \times p} = \begin{pmatrix} \boldsymbol{\Sigma}_{11q \times q} & \boldsymbol{\Sigma}_{12q \times (p-q)} \\ \boldsymbol{\Sigma}_{21(p-q) \times q} & \boldsymbol{\Sigma}_{22(p-q) \times (p-q)} \end{pmatrix}$$

Entonces $\mathbf{x}_1$ está distribuida como $N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

3. Covarianzas cero implica que las correspondientes componentes son independientes. Vale la pena resaltar que esta propiedad solo se cumple en condiciones de normalidad.
4. Las distribuciones marginales de las componentes de $\mathbf{x}$ son normales.
5. Las distribuciones condicionales de las componentes de $\mathbf{x}$ son normales (multivariadas, para $p > 2$).
6. Dada una muestra aleatoria de n observaciones p -variadas, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ donde las $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, los estimadores de máxima verosimilitud de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son, respectivamente, $\bar{\mathbf{x}}$ (ver ecuación (4.5)) y $\mathbf{S}_n$ (ver ecuación (4.6)).
7. $(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2$, dado que si se hace $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu})$, entonces $(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{z}^t \mathbf{z} = \sum_{j=1}^p z_j^2$, donde los $z_j \sim N(0, 1)$, e independientes, y $z_i^2 \sim \chi_1^2$; por tanto, la suma de p variables chi cuadrados independientes con 1 grado de libertad, es chi cuadrado con p grados de libertad.

4.4 Distancia estadística *versus* distancia euclidiana

4.4.1 Distancia euclidiana

Se le conoce también como distancia de línea recta. Considere un caso bidimensional con variables $\mathbf{x}_1$ y $\mathbf{x}_2$ y el vector de medias $\boldsymbol{\mu}^t = (\mu_1, \mu_2)$;

sea el par de observaciones (x_1, x_2) , entonces la distancia euclidiana de este punto al vector de medias es:

$$D_{euc}^2 = [(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2] \quad (4.44)$$

esta cantidad no da cuenta ni de la variabilidad de X_1 y X_2 ni de la covarianza entre ellas. Si se representa gráficamente la distancia para todos los puntos que están ubicados a la misma distancia del vector de medias, resulta una circunferencia, como muestra la figura 4.5.

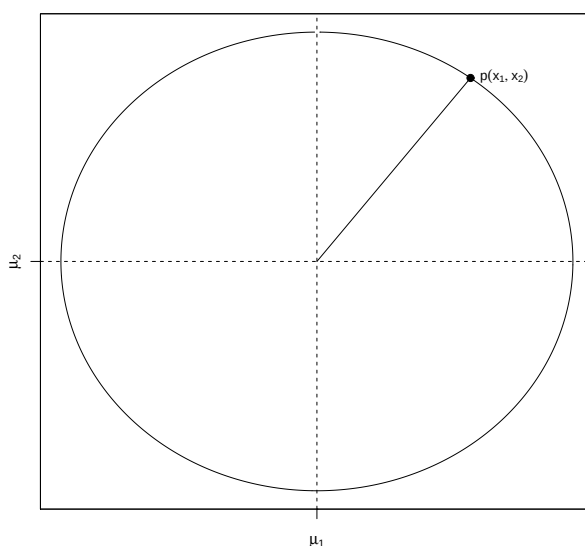


FIGURA 4.5 Representación geométrica de la distancia euclidiana.

4.4.2 Distancia estadística

Una medida que da cuenta de las relaciones lineales entre las variables y su variabilidad es la distancia estadística o Distancia de Mahalanobis respecto al vector de medias (ver la ecuación (4.37)). La distancia de Mahalanobis de la i -ésima observación al vector de medias muestral está dada por:

$$SD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^t \mathbf{S}_u^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (4.45)$$

Para el caso bivariado, con una muestra de n observaciones, se tiene $\bar{\mathbf{x}}^t = (\bar{x}_1, \bar{x}_2)$ y

$$\mathbf{S}_u = \begin{pmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{pmatrix}$$

La covarianza muestral entre X_1 y X_2

$$S_{12} = \left(\frac{1}{n-1} \right) \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \quad (4.46)$$

La varianza muestral de las n observaciones de la variable X_1

$$S_1^2 = \left(\frac{1}{n-1} \right) \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 \quad (4.47)$$

La varianza muestral de la variable X_2

$$S_2^2 = \left(\frac{1}{n-1} \right) \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 \quad (4.48)$$

entonces

$$SD_i^2 = \frac{1}{1 - r_{12}^2} \left[\frac{(x_1 - \bar{x}_1)^2}{S_1^2} - \frac{2r_{12}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{S_1 S_2} + \frac{(x_2 - \bar{x}_2)^2}{S_2^2} \right] \quad (4.49)$$

donde $r_{12} = S_{12}/S_1 S_2$ es el coeficiente de correlación muestral entre X_1 y X_2 . Suponiendo que el par de variables se distribuyen conjuntamente en forma normal bivariada, geométricamente, la región definida por todos los puntos que poseen la misma distancia estadística es una elipse, como la que aparece en la figura 4.6, donde la configuración asume correlación positiva.

Cuando las dos variables son independientes y de varianza igual, la figura será una circunferencia, como en la figura 4.5. El control estadístico multivariado está basado en métodos que usan el concepto de la distancia estadística.

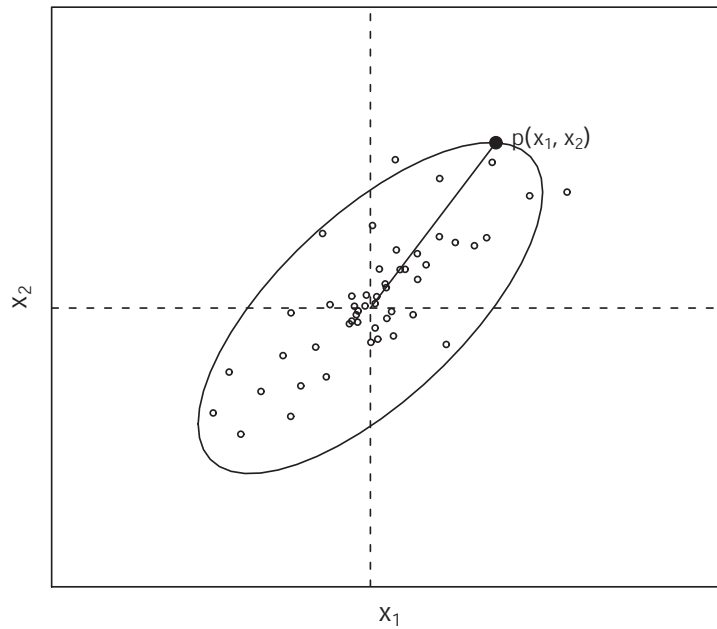


FIGURA 4.6 Representación geométrica de la distancia estadística de dos variables con distribución normal bivariada, coeficiente de correlación positivo.

4.5 Estadístico T^2 de Hotelling

Para el caso univariado, se sabe que el estadístico t , definido como

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \quad (4.50)$$

se distribuye como una $t_{(n-1)}$.

Ahora bien,

$$t^2 = \frac{(\bar{x} - \mu)^2}{s^2/n} = n(\bar{x} - \mu)(s^2)^{-1}(\bar{x} - \mu) \quad (4.51)$$

Hotelling (1931) extendió el estadístico univariado t^2 al caso multivariado, usando estimaciones muestrales de la matriz de varianzas covarianzas. Su derivación se describe a continuación.

Considere una muestra de n observaciones, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde $x_{i(1 \times p)}$ contiene las p medidas de la i -ésima unidad muestral. El estadístico T^2 está dado por

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^t \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \quad (4.52)$$

y se le llama T^2 de Hotelling.

4.5.1 Algunas propiedades

1. El estadístico T^2 es invariante bajo transformaciones lineales no singulares: $\mathbf{x} \rightarrow \mathbf{A}\mathbf{x} + \mathbf{b}$.
2. Sean $\bar{\mathbf{x}}$ y $\mathbf{S}_u$ el vector de medias y la matriz de varianzas covarianzas, respectivamente, de una muestra aleatoria de tamaño n de una $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; entonces:

$$(\bar{\mathbf{x}} - \boldsymbol{\mu})^t \mathbf{S}_u^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim \frac{p(n-1)}{n(n-p)} F_{p, n-p} \quad (4.53)$$

4.5.2 Otros resultados importantes

Sea $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ una muestra aleatoria de n observaciones de una distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Entonces:

1. $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$. Esto se deriva considerando a la media muestral como una combinación lineal dada por la ecuación (4.42), con $a_i = 1/n$.
2. $(n-1)\mathbf{S}_u$ se distribuye como una matriz aleatoria Wishart con $n-1$ grados de libertad. Para detalles de la distribución Wishart, ver Johnson & Wichern (1998).
3. $\bar{\mathbf{x}}$ y $\mathbf{S}_u$ son independientes.

Suponga que $\mathbf{x}_1, \dots, \mathbf{x}_n$ son observaciones independientes de una población con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas covarianzas $\boldsymbol{\Sigma}$ no singular; entonces:

Consecuencias de la ley de los grandes números

Por la ley de los grandes números, cada $\bar{x}_i$ converge en probabilidad a μ_i , $i = 1, 2, \dots, p$; por tanto $\bar{\mathbf{x}}$ converge en probabilidad a $\boldsymbol{\mu}$. También, cada covarianza muestral S_{ij} converge en probabilidad a σ_{ij} , $i, j = 1, 2, \dots, p$, y por consiguiente $\mathbf{S}_n$ (o $\mathbf{S}_u$, ver ecuación (4.10)) converge en probabilidad a $\boldsymbol{\Sigma}$.

Teorema de límite central

Para muestras grandes,

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \stackrel{approx.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Sigma}) \quad (4.54)$$

Para $n - p$ grande,

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})^t \mathbf{S}_u^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \stackrel{approx.}{\sim} \chi_p^2 \quad (4.55)$$

4.6 Evaluación del supuesto de normalidad multivariada

En el contexto multivariado se puede demostrar que si la distribución conjunta es normal multivariada, entonces las distribuciones marginales son normales. Sin embargo, si las distribuciones marginales son normales, esto no implica que la distribución conjunta resulte normal multivariada. Con base en lo anterior, se puede decir que una condición necesaria –mas no suficiente– para la normalidad multivariada es la normalidad univariada de cada una de las variables del problema considerado, pero si alguna de estas distribuciones marginales no es normal, entonces la distribución conjunta tampoco lo es. Por esta razón, un paso inicial en el chequeo de la multinormalidad es determinar si las distribuciones marginales son normales; si esto es así, se puede aplicar cualquiera de los métodos disponibles para probar multinormalidad.

Debido a la gran dispersión inherente de los datos multivariados, las pruebas para normalidad multivariada son poco potentes; sin embargo,

algunos chequeos sobre la distribución son deseables (Rencher, 1995). Se han desarrollado muchos procedimientos para establecer la normalidad multivariada; a continuación se presentan tres de ellos (Rencher, 1995).

4.6.1 Procedimiento 1

Se basa en la realización de gráficos en una y dos dimensiones. En una variable, el supuesto de normalidad multivariada lleva a normalidad univariada, lo que se puede chequear con histogramas de frecuencias o con *boxplots*. En dos variables, el supuesto de normalidad se puede chequear a través de gráficos de dispersión, al notar que estos deberían ser contornos elípticos de densidad constante. Si p no es muy grande, los gráficos de dispersión de cada par de variables se pueden presentar en una matriz de dispersión. A continuación un ejemplo en R, en el cual se simulan 500 observaciones de una normal multivariada con

$$\mu = \begin{pmatrix} 28,100 \\ 7,180 \\ 3,089 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 140,54 & 49,68 & 1,94 \\ 49,68 & 72,25 & 3,68 \\ 1,94 & 3,68 & 0,25 \end{pmatrix}$$

Código R 4.2

```
#Simulando 500 observaciones de una normal trivariada
library(MASS)
mu<-c(28.1,7.18,3.089)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,
0.25),ncol=3,
byrow=T)
datos.simulados<-mvrnorm(500,mu=mu,Sigma=Sigma)

#Función para matriz de dispersión con histogramas:
panel.hist <- function(x, ...){
  usr <- par("usr"); on.exit(par(usr))
  #para definir región de graficación
  par(usr = c(usr[1:2], 0, 1.5) )
  #para obtener una lista que guarde las
  #marcas de clase y conteos en cada una:
  h <- hist(x, plot = FALSE); breaks <- h$breaks;
  nB <- length(breaks); y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
```

```
#para dibujar los histogramas
}

#Aplicando la función panel.hist a datos simulados
pairs(datos.simulados, panel=panel.smooth, cex = 1.5,
pch = 19, bg="light blue",
diag.panel=panel.hist, cex.labels = 1, font.labels=1)
par(oma=c(1,1,1,1),new=T,font=2,cex=0.5)
mtext(outer=T,"Matriz de Dispersión con Histogramas",side=3)

#Función para matriz de dispersión con boxplots:
panel.box <- function(x, ...){
usr <- par("usr",bty='n'); on.exit(par(usr))
par(usr = c(-1,1, min(x)-0.5, max(x)+0.5))
b<-boxplot(x,plot=FALSE); whisker.i<-b$stats[1,]
whisker.s<-b$stats[5,]; hinge.i<-b$stats[2,]
mediana<-b$stats[3,]; hinge.s<-b$stats[4,]
rect(-0.5, hinge.i, 0.5,mediana,...,col='grey')
segments(0,hinge.i,0,whisker.i,lty=2)
segments(-0.1,whisker.i,0.1,whisker.i)
rect(-0.5, mediana, 0.5,hinge.s,...,col='grey')
segments(0,hinge.s,0,whisker.s,lty=2)
segments(-0.1,whisker.s,0.1,whisker.s)
}

#Aplicando la función panel.box a datos simulados
pairs(datos.simulados,panel=panel.smooth,cex = 1, pch = 19,
bg="light blue",diag.panel=panel.box,cex.labels = 0.8,
font.labels=0.8)
par(oma=c(1,1,1,1),new=T,font=2,cex=0.5)
mtext(outer=T,"Matriz de Dispersión con Boxplots",side=3)
```

Las gráficas resultantes aparecen en las figuras 4.7 y 4.8. Los gráficos bivariados de cada par de variables permiten chequear la existencia de tendencias no lineales, presencia de *outliers* y otros indicios de no normalidad. Recuerde que los subconjuntos de variables de una distribución normal multivariada son también normales multivariadas; en particular, los subconjuntos de pares de variables se distribuyen normales bivariados, luego, en los gráficos de dispersión se espera observar un ajuste en línea recta a la nube de puntos.

Salida R 4.2

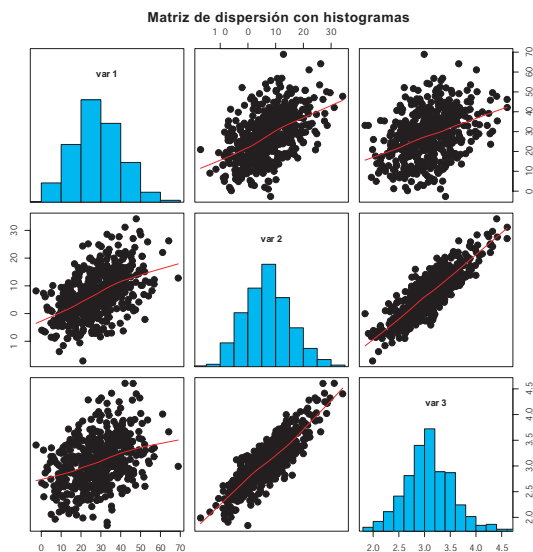


FIGURA 4.7 Matriz de dispersión con histogramas. Se pueden evaluar las características univariadas y las relaciones por pares entre variables.

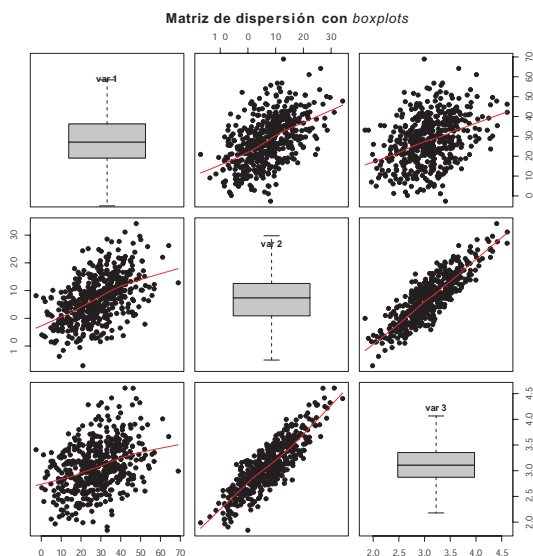


FIGURA 4.8 Matriz de dispersión con *boxplots*. Se pueden evaluar las características univariadas de simetría y dispersión, y las relaciones por pares entre variables.

4.6.2 Procedimiento 2

Esta prueba está basada en asimetría y kurtosis (Mardia, 1970). Sean $\mathbf{x}$ y $\mathbf{y}$ dos vectores aleatorios independientes e idénticamente distribuidos con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas covarianzas $\boldsymbol{\Sigma}$. La asimetría y la kurtosis multivariadas están dadas, respectivamente, por:

$$\begin{aligned}\beta_{1,p} &= E \left[(\mathbf{y} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]^3 \\ \beta_{2,p} &= E \left[(\mathbf{y} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right]^2\end{aligned}\tag{4.56}$$

Ahora bien, dado que los momentos centrales de orden 3 para una distribución normal multivariada son cero, $\beta_{1,p} = 0$. También se puede demostrar que $\beta_{2,p} = p(p+2)$ cuando $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma)$.

Sea

$$g_{ij} = (\mathbf{x}_i - \bar{\mathbf{x}})^t \mathbf{S}_n^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})\tag{4.57}$$

entonces los estimadores de los coeficientes de asimetría y kurtosis con base en una muestra aleatoria de una normal multivariada dada de tamaño n son, respectivamente:

$$\begin{aligned}b_{1,p} &= \hat{\beta}_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n g_{ij}^3 \\ b_{2,p} &= \hat{\beta}_{2,p} = \frac{1}{n} \sum_{i=1}^n g_{ii}^2\end{aligned}\tag{4.58}$$

La tabla A.5 de Rencher (1995) da algunos puntos porcentuales de $b_{1,p}$ y $b_{2,p}$, para $p = 2, 3$ y 4 . Para otros valores de p o cuando $n \geq 50$ se puede recurrir a las siguientes dos pruebas aproximadas:

Para $b_{1,p}$,

$$\begin{aligned}Z_1 &= \frac{(p+1)(n+1)(n+3)}{6[(n+1)(p+1)-6]} b_{1,p} \stackrel{approx.}{\sim} \chi_\nu^2 \\ \nu &= \frac{1}{6} p(p+1)(p+2)\end{aligned}\tag{4.59}$$

rechazar normalidad multivariada si $Z_1 \geq \chi_{0.05, \nu}^2$.

Respecto al coeficiente de kurtosis $b_{2,p}$, se desea rechazar para valores grandes y para valores pequeños. Para los puntos porcentuales superiores de 2,5 % se tiene:

$$Z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \stackrel{aprox.}{\sim} N(0,1) \quad (4.60)$$

Para los puntos porcentuales inferiores de 2,5 %, se tienen dos casos:

1. Si $50 \leq n \leq 400$ se cumple que

$$Z_3 = \frac{b_{2,p} - p(p+2)(n+p+1)/n}{\sqrt{8p(p+2)/(n-1)}} \stackrel{aprox.}{\sim} N(0,1) \quad (4.61)$$

2. Si $n \geq 400$, usar Z_2 .

En resumen, se trata de dos pruebas de hipótesis, una para asimetría y otra para kurtosis.

- Prueba de asimetría: $H_0 : \beta_{1,p} = 0$
- Prueba de kurtosis: $H_0 : \beta_{2,p} = p(p+2)$

Se sabe que si se aceptan las hipótesis nulas en ambas pruebas, Mardia *et al.* (1979) probaron que hay robustez para el estadístico T^2 ; es decir, toda la inferencia clásica con el supuesto de normalidad se puede emplear.

A continuación, la aplicación de la prueba de Mardia usando el paquete R:

Ejemplo 1, datos simulados

Se van a generar $n = 100$ observaciones de una normal multivariada de $p = 3$, que tiene vector de medias $\mu^t = (28, 1, 7, 18, 3, 089)$ y matriz de varianzas covarianzas dada por

$$\Sigma = \begin{pmatrix} 140,54 & 49,68 & 1,94 \\ 49,68 & 72,25 & 3,68 \\ 1,94 & 3,68 & 0,25 \end{pmatrix}$$

La instrucción R correspondiente para generar la matriz de datos de 100×3 es:

Código R 4.3

```
library(MASS)
mu<-c(28.1, 7.18, 3.089)
Sigma<-matrix(c(140.54, 49.68, 1.94, 49.68, 72.25, 3.68, 1.94, 3.68,
0.25),
ncol=3, byrow=T)
X<-mvrnorm(100, mu=mu, Sigma=Sigma)
```

Se va a crear a continuación la función de nombre **Mardia**, con la cual se obtienen los coeficientes muestrales de asimetría y kurtosis, dados por las ecuaciones en (4.58). También calcula los valores críticos con las aproximaciones en las ecuaciones (4.59) a (4.61), usando por defecto un nivel de significancia del 5%. Se ejemplifica su uso sobre la matriz de datos simulados **X**:

Código R 4.4

```
Mardia<-function(X, alpha=0.05){
n<-nrow(X); p<-ncol(X)
Xmedia<-apply(X, 2, mean)
Sn<-((n-1)/n)*var(X)
matriz.media<-matrix(rep(Xmedia, n), ncol=p, byrow=T)
G<-(X-matriz.media)%*%solve(Sn)%*%t(X-matriz.media)
gii<-diag(G); gii2<-gii^2
b2.p<-mean(gii2); G3<-G*G*G
b1.p<-(1/n^2)*sum(G3)
estad.asim<-(p+1)*(n+1)*(n+3)/(6*((n+1)*(p+1)-6))*b1.p
df<-(1/6)*p*(p+1)*(p+2)
Valor.crit.asimet<-qchisq(alpha, df, ncp=0, lower.tail=FALSE,
log.p=FALSE)
Z2<-(b2.p-p*(p+2))/sqrt(8*p*(p+2)/n)
Z3<-(b2.p-p*(p+2)*(n+p+1)/n)/sqrt(8*p*(p+2)/(n-1))
estad.kurt.sup<-Z2
estad.kurt.inf<-c(NA)
estad.kurt.inf[n<50]<-NA
estad.kurt.inf[n>=50 & n<400]<-Z3
estad.kurt.inf[n>400]<-Z2
area<-alpha/2
valor.crit.kurtos1<-qnorm(area, lower.tail = TRUE)
valor.crit.kurtos2<-qnorm(area, lower.tail = FALSE)
resultados<-list(asimetria=b1.p, estad.asim=estad.asim,
Vr.crit.asimet=Valor.crit.asimet, kurtosis=b2.p,
estad.kurt.inf=estad.kurt.inf, estad.kurt.sup=estad.kurt.sup,
vr.crit.kurtos1=valor.crit.kurtos1, vr.crit.kurtos2=valor.crit.
kurtos2)
(resultados<-unlist(resultados))}
```

Al aplicar esta función a la matriz de datos simulada X produce la siguiente salida:

Salida R 4.3

```
Mardia(X)
      asimetria      estad.asim Vr.crit.asimet kurtosis      estad.kurt.inf
      0.6501128      11.3285154   18.3070381   14.2870609   -1.1925361

estad.kurt.sup vr.crit.kurtos1 vr.crit.kurtos2
      -0.6508213      -1.9599640      1.9599640
```

Los valores observados de los estadísticos de prueba para la simetría y la kurtosis son, respectivamente: $Z_1 = 11,3285154$, $Z_2 = -0,6508213$ y $Z_3 = -1,1925361$ (recuerde que para $n = 100$ la kurtosis tiene dos estadísticos). Las regiones críticas al nivel de significancia dado, para los estadísticos de prueba para la asimetría y la kurtosis, son: $Z_1 > 18,3070381$, $Z_3 < -1,9599640$ y $Z_2 > 1,9599640$; por tanto, comparando los valores observados con los críticos, se concluye que la hipótesis de multinormalidad se puede aceptar.

Ejemplo 2, con pesos de corcho

A continuación se presentan los datos referentes a los pesos, en centigramos, del corcho hallado en muestras tomadas en las direcciones norte (N), este (E), oeste (O) y sur (S) del tronco de 28 árboles cultivados en una parcela experimental. Las variables corresponden a los pesos en cada dirección.

Se introducen en R estos datos en forma matricial y se aplica la función `Mardia`, como se muestra a continuación:

Código R 4.5

```
datos<-rbind(c(72,66,76,77),c(60,53,66,63),c(56,57,64,58),
c(41,29,36,38),c(32,32,35,36),c(30,35,34,26),c(39,39,31,27),
c(42,43,31,25),c(37,40,31,25),c(33,29,27,36),c(32,30,34,28),
c(63,45,74,63),c(54,46,60,52),c(47,51,52,43),c(91,79,100,75),
c(56,68,47,50),c(79,65,70,61),c(81,80,68,58),c(78,55,67,60),
c(46,38,37,38),c(39,35,34,37),c(32,30,30,32),c(60,50,67,54),
c(35,37,48,39),c(39,36,39,31),c(50,34,37,40),c(43,37,39,50),
c(48,54,57,43))
```

```
datos; Mardia(datos)
```

TABLA 4.2 Pesos de corcho.

Obs	N	E	S	O	Obs.	N	E	S	O
1	72	66	76	77	15	91	79	100	75
2	60	53	66	63	16	56	68	47	50
3	56	57	64	58	17	79	65	70	61
4	41	29	36	38	18	81	80	68	58
5	32	32	35	36	19	78	55	67	60
6	30	35	34	26	20	46	38	37	38
7	39	39	31	27	21	39	35	34	37
8	42	43	31	25	22	32	30	30	32
9	37	40	31	25	23	60	50	67	54
10	33	29	27	36	24	35	37	48	39
11	32	30	34	28	25	39	36	39	31
12	63	45	74	63	26	50	34	37	40
13	54	46	60	52	27	43	37	39	50
14	47	51	52	43	28	48	54	57	43

Fuente: Díaz (2002), p. 102.

Salida R 4.4

```

asimetria      estad.asim    Vr.crit.asimet    kurtosis      estad.kurt.inf
4.4763816      24.1263013      31.4104328      22.9568700      NA

estad.kurt.sup    vr.crit.kurtos1    vr.crit.kurtos2
-0.3983518      -1.9599640      1.9599640

```

En los anteriores resultados se puede observar que no hay valor crítico para el estadístico inferior de la kurtosis. Los otros valores críticos que aparecen no son confiables puesto que las aproximaciones no funcionan bien en muestras pequeñas ($n = 28 < 50$). ¿Qué hacer? Los valores en la tabla A.5 de Rencher (1995) son limitados, pero por simulación es posible, para un tamaño de muestra y dimensión dados, hallar los valores críticos de la asimetría y kurtosis de la distribución normal p -variada correspondiente. A continuación se presenta un programa en R que tiene tal utilidad, con dos aplicaciones:

Código R 4.6

```

Mardia3<-function(X,n,p){
  Sn<-((n-1)/n)*var(X)
  Xmedia<-apply(X, 2, mean)
  matriz.media<-matrix(rep(Xmedia,n),ncol=p,byrow=T)
  G<-(X-matriz.media)%*%solve(Sn)%*%t(X-matriz.media)
  gii<-diag(G)
}

```

```

gii2<-gii^2
b2.p<-mean(gii2)
G3<-G*G*G
b1.p<-(1/n^2)*sum(G3)
(resultados<-cbind(b1.p,b2.p))}

simul<-function(n,p,Nsimul,alpha){
  library(MASS)
  mu<-c(rep(0,p))
  Sigma<-diag(1,ncol=p,nrow=p)
  a<-matrix(rep(n,Nsimul),ncol=Nsimul)
  X<-array(apply(a,2,mvrnorm,mu,Sigma),dim=c(n,p,Nsimul))
  res<-matrix(apply(X,c(3),Mardia3,n=n,p=p),byrow=T,ncol=2)
  res<-data.frame(list(b1.p=res[,1],b2.p=res[,2]))
  area<-1-alpha
  area1<-alpha/2
  area2<-1-alpha/2
  Kurtos.inf<-quantile(res[,2],probs=area1)
  Kurtos.sup<-quantile(res[,2],probs=area2)
  asim.sup<-quantile(res[,1],probs=area)
  unlist(list(Kurtos.inf=Kurtos.inf,
  Kurtos.sup=Kurtos.sup,asim.sup=asim.sup))}

```

En el anterior código R, la función `Mardia3` calcula sobre una muestra p -variada de tamaño n los respectivos coeficientes de asimetría y kurtosis, en tanto que la función `simul` aproxima mediante simulación Monte Carlo los valores críticos para tales coeficientes usando su distribución empírica resultante al aplicar la función `Mardia3` en cada una de N_{simul} muestras simuladas de una normal p -variada $N_p(\mathbf{0}, \mathbf{I})$ de tamaño n .

Salida R 4.5

```

valores.crit<-simul(n=100,p=3,Nsimul=10000,alpha=0.05)
valores.crit

```

Kurtos.inf.2.5%	Kurtos.sup.97.5%	asim.sup.95%
13.036147	16.892567	1.091322

```

valores.crit<-simul(n=28,p=4,Nsimul=10000,alpha=0.05)
valores.crit

```

Kurtos.inf.2.5%	Kurtos.sup.97.5%	asim.sup.95%
19.445199	26.510923	6.111786

Observe que en el anterior programa se utilizó un número grande de simulaciones: 10000. Retome el ejemplo 1: a un nivel del 5 %, $n = 100$ y $p = 3$, se rechaza la hipótesis de multinormalidad si $b_{1,p} > 1,091322$

o si $b_{2,p} > 16,892567$, o $b_{2,p} < 13,036147$, y comparando con los valores observados se llega a la conclusión de no rechazo de la hipótesis de multinormalidad. Para el ejemplo 2: a un nivel del 5%, $n = 28$ y $p = 4$, se rechaza la hipótesis de multinormalidad si $b_{1,p} > 6,111786$ o si $b_{2,p} > 26,510923$, o $b_{2,p} < 19,445199$, y comparando con los valores observados se concluye que la hipótesis de multinormalidad no se rechaza. A continuación aparecen los programas en SAS IML para el cálculo de asimetría y kurtosis, y para la determinación de los valores críticos empíricos.

Código SAS 4.1

```
START MARDIA(X);
N=NROW(X);
P=NCOL(X);
U=REPEAT(1,n,1);
XB=X'*U/n;
IF P>1 THEN DO;
S=(n/(n-1))*(((1/n)*(X'*X))-(XB*XB'));
A=(X-U*XB')*inv(s)*(X-U*XB')';
B1P=(1/(N**2))*SUM(A##3); /*INDICE DE ASIMETRIA MULTIVARIADA*/
B2P=(1/N)*SUM(VECDIAG(A)##2); /*INDICE DE KURTOSIS MULTIVARIADA*/
END;
IF P=1 THEN DO;
U=REPEAT(1,N,1);
S=(X'*X)/n-XB*XB';
H=X-XB*U;
B1P=(1/N)*SUM(H##3)*(S**(-3/2)); /*INDICE DE ASIMETRIA UNIVARIADA*/
B2P=(1/N)*SUM(H##4)*(S**(-2)); /*INDICE DE KURTOSIS UNIVARIADA*/
END;
print B1P B2P;
FINISH MARDIA;
use M1;
read all into y;
run mardia(y);
quit;
```

4.6.3 Otro procedimiento: gráfico Q-Q chi cuadrado

Con una distribución normal multivariada $Np(\mu, \Sigma)$, y muestras aleatorias con n y $n - p$ grandes, la distancia estadística o distancia de Mahalanobis SD_i^2 (ver ecuación (4.45)) se distribuye aproximadamente como una chi cuadrado con p grados de libertad.

Suponga una muestra aleatoria $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$; las correspondientes distancias $SD_1^2, SD_2^2, \dots, SD_n^2$ se comportan aproximadamente como

una variable con distribución chi cuadrado, a pesar de que no son independientes (Johnson & Wichern, 1998). Los pasos para elaborar el gráfico son los siguientes:

1. Calcular las n distancias SD_i^2 y ordenarlas en forma ascendente, obteniendo la muestra ordenada $SD_{(1)}^2, SD_{(2)}^2, \dots, SD_{(n)}^2$.
2. Graficar los pares $\left(U_i\left(\frac{i-0,5}{n}\right), SD_{(i)}^2\right)$, donde $U_i\left(\frac{i-0,5}{n}\right)$ es el percentil $100\left(\frac{i-0,5}{n}\right)\%$ de la distribución χ_p^2 .
3. Calcular el coeficiente de correlación de Pearson para el gráfico, con el cual se prueba

$$H_0 : \text{corr}\left(U_i\left(\frac{i-0,5}{n}\right), SD_{(i)}^2\right) = 0$$

$$H_1 : \text{corr}\left(U_i\left(\frac{i-0,5}{n}\right), SD_{(i)}^2\right) \neq 0$$

que es una forma de demostrar asociación lineal, junto con el análisis visual del gráfico para establecer si los puntos se ajustan a una línea recta con pendiente de 45° .

A continuación se presenta una función creada en R, con el nombre de `graficochi2`, para realizar tanto el gráfico $Q-Q$ chi cuadrado como la prueba de Pearson. Por defecto la función realiza el test de Pearson al 5 %:

Código R 4.7

```
graficochi2<-function(X, alpha=0.05){
  n<-nrow(X)
  p<-ncol(X)
  Xmedia<-apply(X,2,mean)
  Sn<-((n-1)/n)*cov(X)
  Mahal<-mahalanobis(X,center=Xmedia,cov=Sn)
  Ui<-qchisq(ppoints(n,a=0.5), df=p)
  qqplot(Ui, Mahal,
    main=expression(
      "Q-Q plot de las distancias de Mahalanobis"* ~D^2 ~" vs.
      cuantiles de" * ~ chi[p]^2),
    cex.main=0.8,cex.lab=0.7,xlab=expression(U[i]))
  abline(0, 1, col = 'gray',lty=2)
  #Prueba de pearson sobre correlación ebtre percentiles
  #teóricos y percentiles observados en el gráfico de
  #probabilidad chi cuadrado:
```

```
#library(ctest)
Mahal.ord<-sort(Mahal)
confianza<-1-alpha
pruebacorr<-cor.test(Ui, Mahal.ord,alternative ="two.sided",
method = "pearson",conf.level = confianza)
pruebacorr
}
```

Aplicando la anterior función a los datos del ejemplo con datos simulados (ver página 95, código R 4.3) usados para ejemplificar el procedimiento 2 (el test de Mardia) para probar multinormalidad, se obtiene:

Salida R 4.6
graficochi2(X)

Pearson's product-moment correlation

```
data:  Ui and Mahal.ord
t = 79.3056, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9885558 0.9948210
sample estimates:
      cor
0.992299
```

En la anterior salida se observa una correlación de 0,992299, el valor del estadístico de prueba $t = 79,3056$, los grados de libertad $n - 2 = 98$ y el valor p de la prueba $p\text{-value} < 2,2 \times 10^{-16}$, con la cual se rechaza la hipótesis de no correlación y se concluye que existe relación lineal entre U_i y $SD_{(i)}^2$; por tanto, se tiene evidencia de que la muestra original puede ser de una normal multivariada. El gráfico chi cuadrado arrojado por la función aparece en la figura 4.9.

Para el chequeo analítico de la significancia del coeficiente de correlación de Pearson, se recomienda también usar los valores empíricos de la tabla T.5, p. 466 del libro de Sharma (1996).

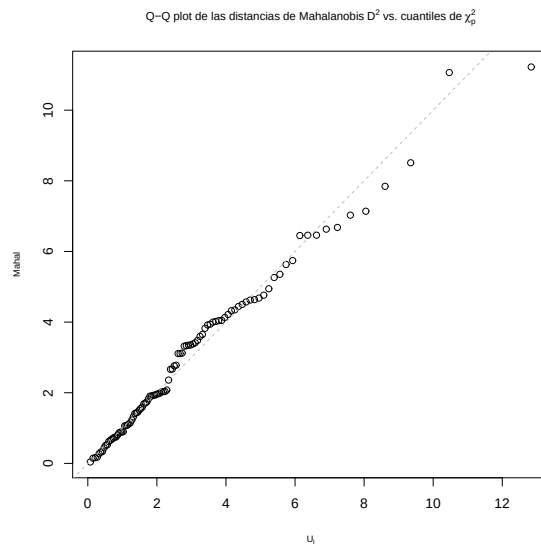


FIGURA 4.9 Gráfico chi cuadrado; datos simulados presentados en el ejemplo con datos simulados del procedimiento 2 para evaluación de normalidad multivariada (página 94).

También se aplicó la función `graficochi2` a los datos de los pesos de corcho. Los resultados son los siguientes:

Salida R 4.7

```
graficochi2(datos)
Pearson's product-moment correlation

data:  Ui and Mahal.ord
t = 32.1812, df = 26, p-value = < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9732107 0.9943554

sample estimates:
      cor
0.9876788
```

donde se obtuvo un coeficiente de correlación de 0.9876788 con un p-value $< 2,2 \times 10^{-16}$, es decir, se rechaza la hipótesis de no correlación y se concluye que existe relación lineal entre U_i y $SD_{(i)}^2$. Por tanto se tiene evidencia de que los datos sobre los pesos del corcho pueden ser de una normal multivariada. El gráfico chi cuadrado arrojado por la función aparece en la figura 4.10.

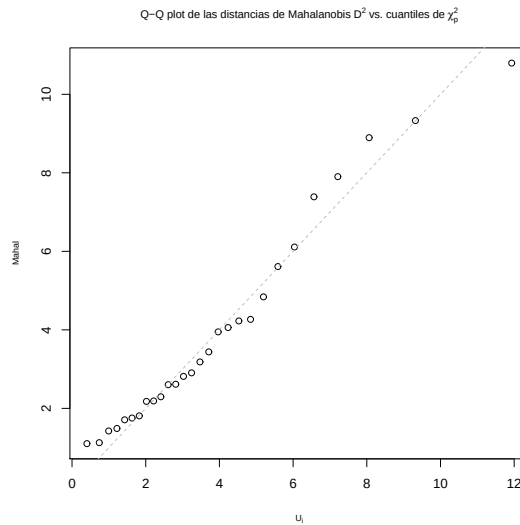


FIGURA 4.10 Gráfico chi cuadrado, datos pesos del corcho, tabla 4.2.

4.7 El estadístico de control T^2 bajo normalidad

Bajo normalidad multivariada, el estadístico dado por

$$SD_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \quad (4.62)$$

es una forma del estadístico T^2 de Hotelling. Se verán otras más dependiendo de los parámetros del proceso que se conozcan y del procedimiento de muestreo empleado en el monitoreo.

4.7.1 Propiedades distribucionales del estadístico T^2 de Hotelling y determinación del límite de control superior (UCL)

Como en el caso de las cartas de control univariada, con base en la distribución del estadístico en consideración T^2 , para el caso multivariado se elige un α y se determina el valor crítico de la distribución. Con las cartas T^2 , se determina solamente el límite superior de control (UCL), pues el límite de control inferior se puede considerar siempre cero (ver figura 4.2).

Se tiene como supuesto básico que las observaciones son una muestra aleatoria de una distribución p normal multivariada $N_p(\mu, \Sigma)$, tomadas de datos históricos del proceso en estado estable de control. El estadístico de control T^2 se puede calcular con base en la media de una muestra de tamaño m (subgrupo racional) tomada durante un periodo de tiempo fijo, o con base en una observación individual de p componentes en un punto muestral fijo. Además, según sean conocidos o no los parámetros del proceso (el vector de medias y la matriz de varianzas covarianzas), resultará una distribución para el estadístico T^2 . A continuación los casos:

μ y Σ conocidos

Suponiendo conocidos los parámetros de la distribución normal multivariada que caracteriza al proceso, para un vector de observaciones individuales $\mathbf{x}$ el estadístico T^2 tiene la forma y la distribución dada por

$$T^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2 \quad (4.63)$$

μ y Σ desconocidos y estimados usando $\bar{\mathbf{x}}$ y $\mathbf{S}_u$

A partir de un conjunto de datos históricos de n vectores de observaciones individuales, se calculan las estimaciones de los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. En el cálculo del estadístico T^2 se tienen los siguientes casos:

■ Un vector de observaciones independiente de los estimadores $\bar{\mathbf{x}}$ y $\mathbf{S}_u$

Si $\mathbf{x}$ es un vector de observaciones individuales, independiente de $\bar{\mathbf{x}}$ y $\mathbf{S}_u$, es decir, no fue usado en el cálculo de estos estimadores, entonces

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{S}_u^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \sim \frac{p(n+1)(n-1)}{n(n-p)} F_{p, n-p} \quad (4.64)$$

donde $F_{p, n-p}$ denota una variable aleatoria con distribución F con p y $(n-p)$ grados de libertad.

- **Un vector de observaciones no independiente de los estimadores $\bar{\mathbf{x}}$ y $\mathbf{S}_u$**

Si $\mathbf{x}$ es un vector de observación no independiente de $\bar{\mathbf{x}}$ y $\mathbf{S}_u$, es decir, fue usado en el cálculo de estos estimadores, entonces

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{S}_u^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \sim \frac{(n-1)^2}{n} B_{p/2, (n-p-1)/2} \quad (4.65)$$

donde $B_{p/2, (n-p-1)/2}$ es una variable aleatoria con distribución beta con parámetros $p/2$ y $(n-p-1)/2$.

Nota 1. Recordar que una variable aleatoria con distribución beta está asociada a una variable aleatoria con distribución F de la siguiente manera: Si $X \sim B(p, q)$, entonces, la variable

$$\frac{p(X^{-1} - 1)}{q} \sim F_{2q, 2p}$$

Luego

$$B_{p/2, (n-p-1)/2} = \frac{pF_{p, n-p-1}}{(n-p-1) + pF_{p, n-p-1}}$$

Nota 2. Los valores T^2 obtenidos con cada una de las observaciones del mismo conjunto de datos históricos ($\bar{\mathbf{x}}$ y $\mathbf{S}_u$ son los mismos) poseen una interdependencia débil,

$$\text{corr}(T_i^2, T_j^2) = -\frac{1}{n-1}$$

luego, si $n \rightarrow \infty$, esta correlación es cero.

μ y Σ desconocidos y estimados usando m subgrupos racionales de tamaño n

Cuando se monitorea la media de muestras de tamaño n , tomadas en cada uno de m intervalos de muestreo, se usan como estimadores de los parámetros, los siguientes:

$$\bar{\bar{\mathbf{x}}} = \frac{1}{m} \sum_{j=1}^m \bar{\mathbf{x}}_j \quad (4.66)$$

$$\bar{\mathbf{S}}^2 = \frac{1}{m} \sum_{j=1}^m \mathbf{S}_{u,j} \quad (4.67)$$

Se tienen dos casos para la distribución del estadístico T^2 :

- $\bar{\mathbf{x}}$ es independiente de $\bar{\bar{\mathbf{x}}}$ y de $\bar{\mathbf{S}}^2$
En este caso, $\bar{\mathbf{x}}$ no fue usado en la construcción de los estimadores $\bar{\bar{\mathbf{x}}}$ y $\bar{\mathbf{S}}^2$; entonces

$$T^2 = (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})^t \bar{\mathbf{S}}^{-1} (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}) \sim \frac{p(m+1)(n-1)}{n(mn-m-p+1)} F_{p, mn-m-p+1} \quad (4.68)$$

- $\bar{\mathbf{x}}$ no es independiente de $\bar{\bar{\mathbf{x}}}$ y de $\bar{\mathbf{S}}^2$
En este caso, $\bar{\mathbf{x}}$ fue usado en la construcción de los estimadores $\bar{\bar{\mathbf{x}}}$ y $\bar{\mathbf{S}}^2$; entonces

$$T^2 = (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})^t \bar{\mathbf{S}}^{-1} (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}) \sim \frac{p(m-1)(n-1)}{n(mn-m-p+1)} F_{p, mn-m-p+1} \quad (4.69)$$

4.8 Chequeo de supuestos para el uso del estadístico T^2

4.8.1 Normalidad multivariada

En la construcción de la carta T^2 se asume que:

- El proceso es normal multivariado
- Las observaciones son independientes

Cualquier violación de estos supuestos puede alterar significativamente la eficiencia del procedimiento. Previamente se presentaron varios métodos para establecer la normalidad multivariada. Otra alternativa se explica a continuación.

Mason & Young (2002) sugieren desarrollar una prueba de bondad de ajuste sobre los valores $C T^2$ para determinar si la distribución beta o F , según el caso (ver sección 4.7.1), se ajustan adecuadamente, donde C es una constante apropiada según distribución resultante bajo normalidad y el esquema de monitoreo aplicado. Por ejemplo, cuando se analizan n observaciones históricas individuales del proceso p -variado con parámetros desconocidos, se tiene la ecuación (4.65), de donde

$$\frac{n}{(n-1)^2} T_i^2 \sim B_{p/2, (n-p-1)/2}$$

Por tanto, se puede construir un gráfico Q-Q de los valores $\frac{n}{(n-1)^2} T_{(i)}^2$ vs. los cuantiles $(i-0,5)/n$ de la distribución $B_{p/2, (n-p-1)/2}$. Se espera que los puntos ajusten a la línea recta de 45° que pasa por el origen. Se puede determinar dicha linealidad también analíticamente; ver, por ejemplo, Rencher (1995, p. 111). Como una alternativa a estas tablas se hace una prueba de significancia del coeficiente de correlación de Pearson, aunque no sea una alternativa muy precisa.

Nota. Hay que tener en cuenta que la presencia de *outliers* distorsiona los resultados de estas pruebas; sin embargo, los gráficos Q-Q pueden permitir identificar las posibles observaciones *outliers*.

A continuación se presenta una función creada en R para generar este gráfico junto con la prueba de correlación de Pearson para el gráfico, y una aplicación a 500 datos simulados.

Código R 4.8

```
graficobeta<-function(X, alpha=0.05){
  n<-nrow(X); p<-ncol(X); Xmedia<-apply(X,2,mean)
  Su<-cov(X); C.Mahal<-(n/(n-1)^2)*(mahalanobis(X,center=Xmedia,
  cov=Su))
  a<-p/2; b<-(n-p-1)/2
  Ui<-qbeta(ppoints(n,a=0.5), shape1=a, shape2=b,
  lower.tail = TRUE)
  qqplot(Ui,C.Mahal,main=expression(
  "Q-Q plot de "*C.Mahal==n*(n-1)^-2*T[i]^2*" vs. cuantiles de"
  "*U==beta[paste(0.5*p~","*~0.5*(n-p-1))]),cex.main=1,
  cex.lab=0.7,
  xlab=expression(U[i]),ylab="C. Mahal")
  abline(0, 1, col = 'gray', lty=2)
  Mahal.ord<-sort(C.Mahal); confianza<-1-alpha
  pruebacorr<-cor.test(Ui, Mahal.ord,alternative = "two.sided",
  method = "pearson",conf.level = confianza)
```

```
pruebacorr}

#Simulación de datos:
library(MASS); mu<-c(28.1,7.18,3.089)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,
0.25),
ncol=3,byrow=T); datos.simulados<-mvrnorm(500,mu=mu,Sigma=Sigma)
```

Salida R 4.8

```
#Aplicación de la función graficobeta:
graficobeta(datos.simulados)
```

Pearson's product-moment correlation

```
data: Ui and Mahal.ord
t = 313.6824, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.996995 0.997885
sample estimates:
      cor
0.997479
```

En la anterior salida, se tiene que para la muestra de 500 observaciones de una normal multivariada con $p = 3$, existe una fuerte asociación lineal entre las observaciones $\frac{n}{(n-1)^2}T_{(i)}^2$ vs. los cuantiles $(i - 0,5)/n$ de la distribución $B_{p/2,(n-p-1)/2}$, luego es evidente un buen ajuste de esta última distribución. El gráfico arrojado por el procedimiento aparece en la figura 4.11.

4.8.2 Transformaciones y aproximaciones no paramétricas

Si se prueba que los datos no son normales multivariados, en algunos casos puede ocurrir que al realizar transformaciones sobre las variables que específicamente presentan no normalidad, el nuevo conjunto de variables se distribuya en forma normal multivariada. El problema es que las transformaciones pueden dificultar las inferencias sobre el proceso. Como alternativa, se cuenta con los métodos no paramétricos o libres de distribución para establecer el límite de control o UCL de la carta T^2 . A continuación se describen brevemente algunos de estos procedimientos.

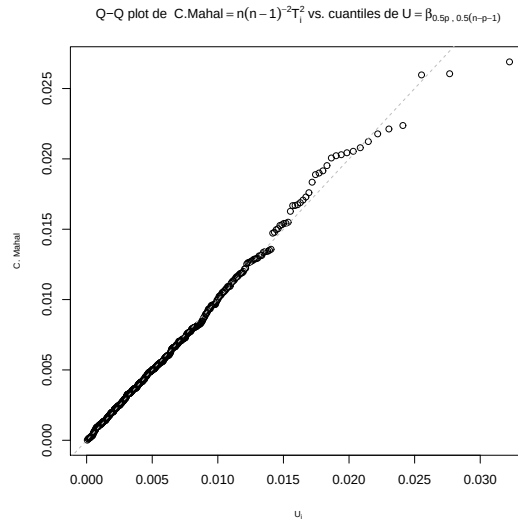


FIGURA 4.11 Gráfico Q-Q Beta para 500 observaciones $\frac{n}{(n-1)^2}T_i^2$ obtenidas por simulación de una distribución normal multivariada con $p = 3$.

Aplicación del teorema de Chebyshev

Cabe recordar que el teorema de Chebyshev establece que para una variable aleatoria X con media μ y varianza σ^2 se cumple que:

$$P(\mu - k\sigma < x < \mu + k\sigma) \geq 1 - \frac{1}{k^2}, \quad \forall k > 0, k \in \mathbb{R}$$

Para el cálculo del UCL de la carta de control T^2 , se obtiene la media $\bar{T}$ y la desviación estándar S_T de los valores T_i^2 del conjunto de datos históricos. Un UCL aproximado está dado por:

$$UCL = \bar{T} + kS_T \quad (4.70)$$

donde el valor de k se determina según el nivel de significancia α deseado (probabilidad de error tipo I). Aplicando Chebyshev, $\alpha = 1/k^2$, de donde $k = \sqrt{\alpha^{-1}}$.

Como se sabe, la desigualdad de Chebyshev es una cota inferior que proporciona resultados conservadores en la medida que puede estar alejada de su valor real.

Intervalo de confianza para el percentil $(1 - \alpha)100\%$ de los valores T^2

Un segundo método para estimar el UCL de un estadístico T^2 se basa en encontrar un límite de confianza para el UCL libre de distribución. Esta aproximación usa el hecho de que el UCL representa el percentil $(1 - \alpha)100\%$ de la distribución de T^2 . Suponga que $T_{(1)}^2, T_{(2)}^2, \dots, T_{(n)}^2$ son los valores muestrales T^2 ordenados (ver Conover, 1998). Sea Q_p el percentil p -ésimo de la distribución de T^2 , con $p = 1 - \alpha$, y α la tasa de falsas alarmas deseada. Un intervalo de confianza del $100\gamma\%$ para $Q_{1-\alpha}$ está dado por:

$$P\left(T_{(r)}^2 \leq Q_{1-\alpha} \leq T_{(s)}^2\right) = \gamma \quad (4.71)$$

donde r y s se hallan solucionando la siguiente desigualdad:

$$\sum_{j=r}^{s-1} \binom{n}{j} (1 - \alpha)^j (\alpha)^{n-j} \geq \gamma \quad (4.72)$$

Puesto que

$$P\left(T_{(r)}^2 \leq Q_p \leq T_{(s)}^2\right) = P\left(Q_p \leq T_{(s)}^2\right) - P\left(Q_p < T_{(r)}^2\right)$$

y

$$\begin{aligned} P\left(Q_p < T_{(r)}^2\right) &= P\left(\text{al menos } n - r + 1 \text{ de los } T_i^2 \text{ son } > Q_p\right) \\ &= P\left(r - 1 \text{ o menos de los } T_i^2 \text{ son } \leq Q_p\right) \\ &\leq \sum_{j=0}^{r-1} \binom{n}{j} p^j (1 - p)^{n-j} \leq (1 - \gamma)/2 \end{aligned}$$

y teniendo en cuenta que los valores T_i^2 tienen una función de distribución continua, también se cumple que:

$$P\left(Q_p \leq T_{(s)}^2\right) \geq \sum_{j=0}^{s-1} \binom{n}{j} p^j (1 - p)^{n-j} \geq 1 - (1 - \gamma)/2$$

Otra manera consiste en la aplicación del método exacto basado en estadísticos de orden de una distribución uniforme $(0,1)$, $U_{(i)}$, teniendo

en cuenta que la función de distribución de T_i^2 , $F(T_i^2)$ es continua:

$$\begin{aligned} P\left(T_{(r)}^2 \leq Q_p \leq T_{(s)}^2\right) &= P\left(F\left(T_{(r)}^2\right) \leq p \leq F\left(T_{(s)}^2\right)\right) \\ &= P\left(U_{(r)} \leq p \leq U_{(s)}\right) \\ &= B_p(r, n-r+1) - B_p(s, n-s+1) \geq \gamma \end{aligned} \quad (4.73)$$

donde

$$B_p(v_1, v_2) = \frac{\Gamma(v_1 + v_2)}{\Gamma(v_1) \Gamma(v_2)} \int_0^p t^{v_1-1} (1-t)^{v_2-1} dt$$

de nuevo, r y s se eligen tan cercanos como sea posible y tal que se alcance el nivel de confianza especificado $100\gamma\%$.

Para n grande, r y s se pueden aproximar como los dos valores dados por:

$$n(1-\alpha) \pm Z_{(1-\gamma)/2} \sqrt{n(1-\alpha)\alpha} \quad (4.74)$$

(4.71) y (4.72) arrojan valores muy similares en muestras grandes y $\gamma \approx 0,95$.

Como UCL se toma a

$$UCL = \frac{T_{(r)}^2 + T_{(s)}^2}{2} \quad (4.75)$$

A continuación se presenta una función en R para aplicar la ecuación (4.72), la cual arroja como resultado los valores de r , s , el nivel de confianza estimado y la diferencia $s-r$; además, se da una aplicación para el intervalo de confianza del percentil 0,99 de una muestra de tamaño $n=300$:

Código R 4.9

```
estima.s.r<-function(alpha,n,gamma=0.95){
p<-1-alpha
conf<-NA
S<-NA
R<-NA
DIF.R.S<-NA
for(r in 1:n){
for(s in 1:n){
if(r<s){
temp1<-pbinom(s-1,n,p)
temp2<-pbinom(r-1,n,p)
aux1.1<-temp1-temp2
conf<-append(conf,aux1.1)
```

```
S<-append(S,s)
R<-append(R,r)
dif<-s-r
DIF.R.S<-append(DIF.R.S,dif)}}}
result<-cbind(conf,S,R,DIF.R.S)
result<-result[-1,]
result<-result[result[,1]>=gamma,]
O<-order(result[,1])
result<-result[O,]
result<-result[1,]
result}
res<-estima.s.r(alpha=0.01,n=300)
```

Salida R 4.9

```
res
      conf      S      R      DIF.R.S
0.9506943 300.0000000 290.0000000 10.0000000
```

En el anterior ejemplo, para una muestra de tamaño 300, el UCL estimado para una tasa de falsa alarma de 0,01 sería el promedio de $T_{(290)}^2$ y $T_{(300)}^2$, y el nivel de confianza para el percentil 0,99 sería de 0,9507 aproximadamente (un poco más del valor esperado de 0,95).

Otra función, también aplicable a la ecuación (4.72), la cual garantiza que $P(Q_p < T_{(r)}^2) \leq (1 - \gamma)/2$ y $P(Q_p \leq T_{(s)}^2) \geq 1 - (1 - \gamma)/2$, se da a continuación:

Código R 4.10

```
sol.r.s<-function(n,alpha,gamma){
  prob<-1-alpha
  r<-qbinom(((1-gamma)/2),size=n,prob=prob)
  s<-qbinom((0.5+gamma/2),size=n,prob=prob)
  conf.correg<-pbinom(s-1,size=n,prob=prob)-pbinom(r-1,size=n,
  prob)
  res<-unlist(list(r=r,s=s,confianza=conf.correg))
  (res<-data.frame(res))}
```

Salida R 4.10

```
sol.r.s(300,0.01,0.95)
      res
r      293.000000
s      300.000000
confianza 0.939485
```

En este caso, para una muestra de tamaño 300, el UCL estimado para una tasa de falsa alarma de 0,01 sería el promedio de $T_{(293)}^2$ y $T_{(300)}^2$, y el

nivel de confianza para el percentil 0,99 sería de aproximadamente 0,939, menor que el nivel esperado de 0,95.

Finalmente, la siguiente función en R aplicaría para usar la ecuación (4.73):

Código R 4.11

```
estima.s.r2<-function(alpha,n,gamma=0.95){
p<-1-alpha
conf<-NA
S<-NA
R<-NA
DIF.R.S<-NA
for(r in 1:n){
for(s in 1:n){
if(r<s){
temp1<-pbeta(p,s,n-s+1)
temp2<-pbeta(p,r,n-r+1)
aux1.1<-temp2-temp1
conf<-append(conf,aux1.1)
S<-append(S,s)
R<-append(R,r)
dif<-s-r
DIF.R.S<-append(DIF.R.S,dif)}}}
result<-cbind(conf,S,R,DIF.R.S)
result<-result[-1,]
result<-result[result[,1]>=gamma,]
O<-order(result[,1])
result<-result[O,]
result<-result[1,]
result}
res<-estima.s.r2(alpha=0.01,n=300)
```

Salida R 4.11

conf	S	R	DIF.R.S
0.9506943	300.0000000	290.0000000	10.0000000

Como se puede ver, los resultados son iguales a los obtenidos con la función `estima.s.r`, debido básicamente al tamaño de muestra.

Con la aproximación normal para el caso del percentil 0,99, con una confianza del 95 % y $n = 300$, los valores de r y s serían, respectivamente:

$$r = 300 \times 0,99 - 1,96 \times \sqrt{300 \times 0,99 \times 0,01} = 293,6 \approx 294$$

$$s = 300 \times 0,99 + 1,96 \times \sqrt{300 \times 0,99 \times 0,01} = 300,4 \approx 300$$

y el nivel de confianza correspondiente

$$\sum_{j=294}^{299} \binom{300}{j} (0,99)^j (0,01)^{300-j} = 0,918 < 0,95$$

UCL aproximado ajustando una distribución al estadístico T^2 mediante suavizamiento kernel

Este método halla el UCL aproximado usando el cuantil $(1 - \alpha)$ de la función de distribución kernel del estadístico T^2 . Este método logra buenas aproximaciones con tamaños de muestra razonablemente grandes ($n > 250$), aunque da estimaciones sesgadas.

La función kernel o distribución kernel del estadístico T^2 está dada por:

$$K(t) = \frac{1}{n} \sum_{j=1}^n \Phi \left\{ \frac{(t - T_j^2)}{h} \right\} \quad (4.76)$$

donde $\Phi(\cdot)$ denota la función normal acumulada estándar, y h es el estimador del ancho de banda o de ventana que determina la cantidad de suavización de la estimación. Específicamente, el método realiza los siguientes pasos:

1. Se ajusta una densidad al conjunto de n valores del estadístico T^2 usando la técnica de suavizamiento kernel; por ejemplo, la función **density** de la librería R **KernSmooth** proporciona m pares (x_i, y_i) , siendo x_i los puntos donde la densidad es estimada y y_i los valores de densidad estimados, satisfaciendo $x_1 < x_2 < \dots < x_m$ y $\sum_{i=1}^m y_i = 1$.
2. Desde la función kernel normal se pueden producir coordenadas x_i negativas, que estarían representando valores T^2 negativos (lo cual no es plausible). A continuación es necesario eliminar los pares (x_i, y_i) tales que $x_i < 0$ y recalcular las densidades para los pares

restantes de manera que las nuevas densidades sumen 1: esto es, para $i : x_i \geq 0$, $\hat{y}_i = y_i / \sum_{i:x_i \geq 0} y_i$.

3. Usando el nuevo conjunto de pares obtenidos en el paso anterior, se halla el cuantil $(1 - \alpha)$ como $\inf\{x_i : \sum_{j=1}^i \tilde{y}_j \geq 1 - \alpha\}$. Este cuantil es el UCL obtenido por el método kernel.

A continuación se presenta el algoritmo en R que calcula el UCL mediante suavizamiento kernel, grafica el histograma de los valores T_i^2 con la densidad kernel superpuesta y estima la tasa de error tipo I esperada con esta estimación:

Código R 4.12

```
UCL.kernel<-function(T2,alpha){
  library(KernSmooth)
  est<-density(T2,kernel="gaussian",bw="nrd")
  aux<-cbind(est$x,est$y); aux<-aux[aux[,1]>=0,]
  estand<-aux[,2]/sum(aux[,2]); n1<-nrow(aux); acum<-0
  for(i in 1:n1){
    if(acum<(1-alpha)){
      j<-i; acum<-sum(estand[1:i])}
    yhist<-hist(T2,plot=FALSE); yhist<-yhist$density
    x.lim<-c(min(est$x),max(est$x,T2)); y.lim<-c(0,max(est$y,yhist))
    hist(T2,freq=FALSE,xlim=x.lim,ylim=y.lim,
         main=expression(paste("Histograma de",sep=" ",T^2,sep=" ",
                                "con suavizamiento kernel")),xlab=expression(T^2),
         ylab="Densidad")
    par(new=T); plot(est,main="",xlab="",ylab="",xlim=x.lim,
                    ylim=y.lim)
    UCL<-aux[j,1]; (res<-unlist(list(UCL=UCL,alpha1=(1-acum))))}
```

Ahora se obtiene por simulación 500 observaciones de una normal multivariada con $p = 3$, se calculan los valores T_i^2 y se estima el UCL para una tasa de error tipo I de 0,01, usando el suavizamiento kernel:

Código R 4.13

```
library(MASS)
mu<-c(28.1,7.18,3.089)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,
0.25),
ncol=3,byrow=T)
X<-mvrnorm(500,mu=mu,Sigma=Sigma);Xmedia<-apply(X,2,mean);
Su<-cov(X)
T2<-mahalanobis(X,center=Xmedia,cov=Su)
```

Salida R 4.12

```
UCL.kernel(T2,0.01)
      UCL      alpha1
11.229759227  0.009987697

quantile(T2,probs=0.99)
      99%
10.51392
```

La gráfica obtenida con la aplicación de la anterior función se presenta en la figura 4.12. Observe que el UCL se estima en 11,229759227, y que la tasa de error tipo I sería $0,009987697 < 0,01$. Compare el UCL estimado con el cuantil 0,99 de la muestra de los T_i^2 .

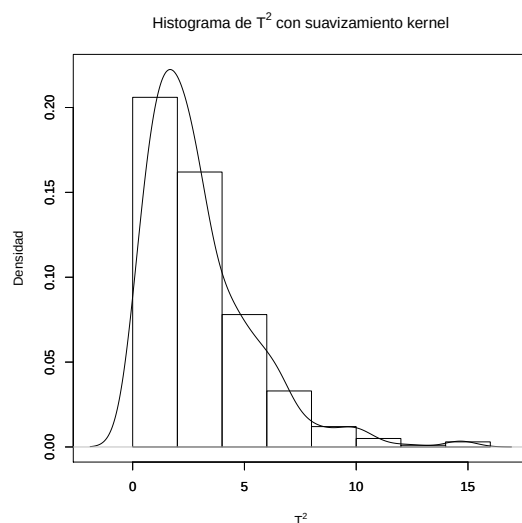


FIGURA 4.12 Histograma de los valores T_i^2 obtenidos de una muestra de 500 observaciones de una normal multivariada con $p = 3$. La curva superpuesta corresponde a la densidad estimada por suavizamiento kernel.

Finalmente, compare el UCL estimado con el verdadero, bajo normalidad (ver μ y Σ desconocidos y estimados usando $\bar{\mathbf{x}}$ y $\mathbf{S}_u$):

$$\begin{aligned} UCL_{exacto} &= \frac{(n-1)^2}{n} B_{0,01,p/2,(n-p-1)/2} \\ &= \frac{(499)^2}{500} B_{0,01,3/2,248} = 11,25020 \end{aligned}$$

4.8.3 Tamaños de muestra

Para alcanzar una precisión dada en las estimaciones de los parámetros del proceso multivariado, el tamaño de muestra debe ser suficientemente grande, de modo que satisfaga el tamaño requerido por cada variable, según las desviaciones estándar. Se podría determinar para cada variable un tamaño de muestra, y luego seleccionar el más grande de todos como el n por usar para el muestreo del proceso multivariado. Este tamaño de muestra debe cumplir ciertas condiciones, que se listan a continuación:

1. n debe exceder a p ; de lo contrario, la matriz de varianza covarianza estimada será singular.
2. n debe ser suficientemente grande comparado con el número total de parámetros por estimar (medias, varianzas y covarianzas), los cuales suman un total de $2p + p(p - 1)/2$, para efectos de obtener estabilidad y exactitud en las estimaciones.
3. n debe ser tan grande como sea necesario, para observar las condiciones de operación normal y para caracterizar los efectos sobre el proceso de factores externos como los ambientales.

Ante un gran número de variables, se podría optar por recurrir a técnicas multivariadas para reducir la dimensionalidad del problema, como cuando se usan componentes principales.

4.9 Construcción de la carta de control T^2

4.9.1 Programación en R

A continuación se presentan cuatro funciones escritas en R, para la obtención de la carta T^2 según la fase de implementación: Fase I o Fase II, y según el esquema de muestreo: observaciones individuales a intervalos fijos, subgrupos racionales de igual tamaño tomados a intervalos fijos.

Carta de control T^2 para la Fase I con n observaciones históricas individuales

En este caso, resulta aplicable la ecuación dada en (4.65). La función programada en R para tal fin es:

Código R 4.14

```
carta.T2.obs.faseI<-function(X,alpha){
  n<-nrow(X)
  p<-ncol(X)
  UCL<-(((n-1)^2)/n)*qbeta((1-alpha),(p/2),(n-p-1)/2)
  Xmedia<-apply(X,2,mean)
  Su<-var(X)
  T2<-mahalanobis(X,center=Xmedia,cov=Su)
  Observacion<-1:n
  par(bg="cornsilk")
  plot(Observacion,T2,type='l',xlim=c(0,n+2),ylim=c(0,max(UCL,
    max(T2))+2),
    main=expression("Carta"*~T^2),
    ylab=expression(T^2),xlab="No. Observación",font=2)
  abline(h=UCL,lty=3)
  for(i in 1:n){
    temp<-ifelse((T2[i]>UCL),4,19)
    points(Observacion[i],T2[i],pch=temp)
    if(T2[i]>UCL)text(i,T2[i],labels=paste('obs=',i),pos=3,font=2,
      cex=0.7) }
    text((max(Observacion)-1),UCL,paste('UCL=',round(UCL,digits=4)),
      pos=3,
      font=2,cex=0.7)
    legend(locator(1),c(paste("p=",p),paste("alpha=",alpha),
      paste("n=",n)),
      ncol=3,cex=0.7,bg='gray95')
    (estimaciones<-list(medias=Xmedia,var=Su,T2=T2))}
```

A continuación se ilustra su uso, sobre un conjunto de datos históricos de tamaño 50, obtenido por simulación, con $p = 3$:

Código R 4.15

```
#Simulación de datos:
library(MASS)
mu<-c(28.1,7.18,3.089)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,
  0.25),
  ncol=3,byrow=T)
X1<-mvrnorm(50,mu=mu,Sigma=Sigma)
#aplicación de la función y resultados obtenidos:
(result<-carta.T2.obs.faseI(X1,0.05))
```

#NOTA: En la ventana gráfica haga clic donde desee colocar la leyenda

Salida R 4.13

```
$medias
[1] 27.712867 6.615687 3.080928

$var
      [,1]      [,2]      [,3]
[1,] 198.234137 71.931956 2.2051096
[2,] 71.931956 81.327921 3.5565836
[3,] 2.205110 3.556584 0.2233516

$T2
[1] 3.2325819 2.1107680 2.5063068 5.5125580 1.9366688 1.1235411 1.9348538
[8] 2.9762154 1.7814212 4.4636039 5.0778266 5.0135528 3.3638625 3.5701256
[15] 7.3553835 3.2232013 5.0081798 0.6590770 0.3837065 3.5694924 1.0110463
[22] 4.6982903 0.7032206 3.0496870 4.6225821 2.0989722 2.1302154 2.8388096
[29] 6.4248152 0.8326738 2.6910077 0.5292737 0.4603333 1.6085409 2.9656233
[36] 5.6508790 1.6410335 8.1263711 2.7447279 0.2946423 0.2274023 3.7146712
[43] 5.0855683 1.7420301 3.3624181 3.3595612 0.9667185 0.6533459 4.9827345
[50] 2.9798779
```

La función `carta.T2.obs.faseI` arroja las estimaciones del vector de medias y de la matriz de varianzas covarianzas, los valores T^2 y el gráfico de la carta, la cual se presenta en la figura 4.13.

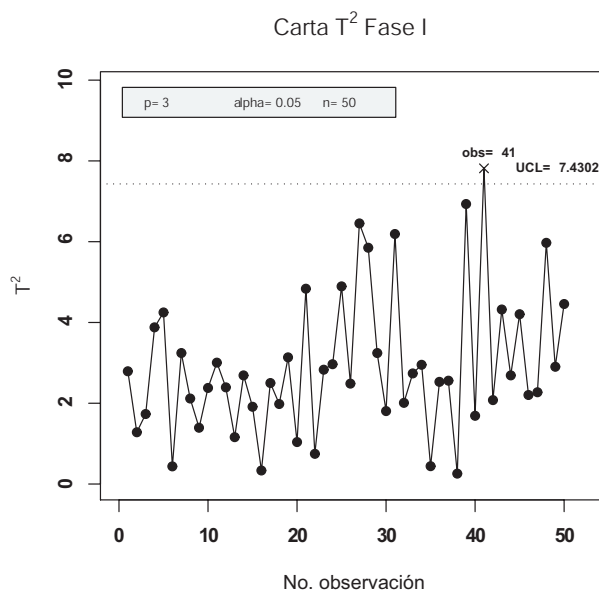


FIGURA 4.13 Carta de control T^2 en la Fase I, con 50 observaciones individuales simuladas. La observación produjo una señal; sin embargo, corresponde a un valor sin causa asignable, originado por la aleatoriedad.

Carta de control T^2 para la Fase II con k observaciones individuales

En este caso, resulta aplicable la ecuación dada en (4.64). La función diseñada en R para su aplicación, se presenta a continuación.

Código R 4.16

```
#Carta T2 para observaciones individuales
#X es el conjunto de nuevas observaciones
#nHDS es el tamaño del conjunto de datos históricos usados
en la Fase I
#medias.est es el vector de medias estimadas en la Fase I
#var.covar.est es la matriz de varianzas covarianzas estimada
en la Fase I

carta.T2.obsII<-function(X,medias.est,var.covar.est,nHDS,
alpha){
  p<-ncol(X)
  k<-nrow(X)
  v1<-p; v2<-nHDS-p
  UCL<-(p*(nHDS+1)*(nHDS-1)/(nHDS*v2))*qf((1-alpha),v1,v2)
  T2<-mahalanobis(X,center=medias.est,cov=var.covar.est)
  Observacion<-1:k; par(bg="cornsilk")
  plot(Observacion,T2,type='l',xlim=c(0,k+2),ylim=c(0,max(UCL,
max(T2))+2),
  main=expression("Carta"~T^2*~"Fase II"),ylab=expression(T^2),
  xlab="No.
Observación",font=2)
  abline(h=UCL,lty=3)
  for(i in 1:k){
    temp<-ifelse((T2[i]>UCL),4,19)
    points(Observacion[i],T2[i],pch=temp)
    if(T2[i]>UCL)text(i,T2[i],labels=paste('obs=',i),pos=3,font=2,
cex=0.7) }
  text((max(Observacion)-1),UCL,paste('UCL=',round(UCL,digits=4)),
pos=3,
font=2,cex=0.7)
  legend(locator(1),c(paste("p=",p),paste("alpha=",alpha),
paste("nHDS=",nHDS),paste('k=',k)),ncol=4,cex=0.7,bg='gray95')
  (resultados<-list(medias=medias.est,var=var.covar.est,T2=T2))
}
```

La función `carta.T2.obsII` proporciona los valores T^2 de las nuevas observaciones, las estimaciones previamente halladas en la Fase I del vector de medias y de la matriz de varianzas covarianzas, y la gráfica de la carta. A continuación se presenta una aplicación a un conjunto

de 20 nuevas observaciones, halladas de nuevo mediante simulación del proceso usado en la Fase I, donde las estimaciones halladas en dicha fase son usadas en la aplicación de la función `carta.T2.obsII`.

Código R 4.17

```
#Simulación de 20 nuevas observaciones:
library(MASS)
mu<-c(28.1,7.18,3.089)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,
0.25),
ncol=3,byrow=T)
X3<-mvrnorm(20,mu=mu,Sigma=Sigma)

#Construcción carta T2 en la Fase II: se usan los parámetros
estimados
#en la Fase I que se suponen guardados en objeto result creado
al
#aplicar la función carta.T2.obs.faseI

(result2<-carta.T2.obsII(X3,result$medias,result$var,50,0.01))

#NOTA: En la ventana gráfica haga click donde desee colocar
la leyenda
```

Salida R 4.14

```
$medias
[1] 27.712867 6.615687 3.080928
$var
      [,1]      [,2]      [,3]
[1,] 198.234137 71.931956 2.2051096
[2,] 71.931956 81.327921 3.5565836
[3,] 2.205110 3.556584 0.2233516
$T2
[1] 1.39589715 1.59767407 0.07934651 0.31366819 3.08875456 1.10410257
[7] 8.64220779 0.42950756 1.56426638 0.74670757 0.69651558 0.41025661
[13] 0.83163871 2.69172603 0.79080041 4.33351965 1.25954082 3.14129558
[19] 6.96731085 1.87144333
```

La gráfica correspondiente se presenta en la figura 4.14.

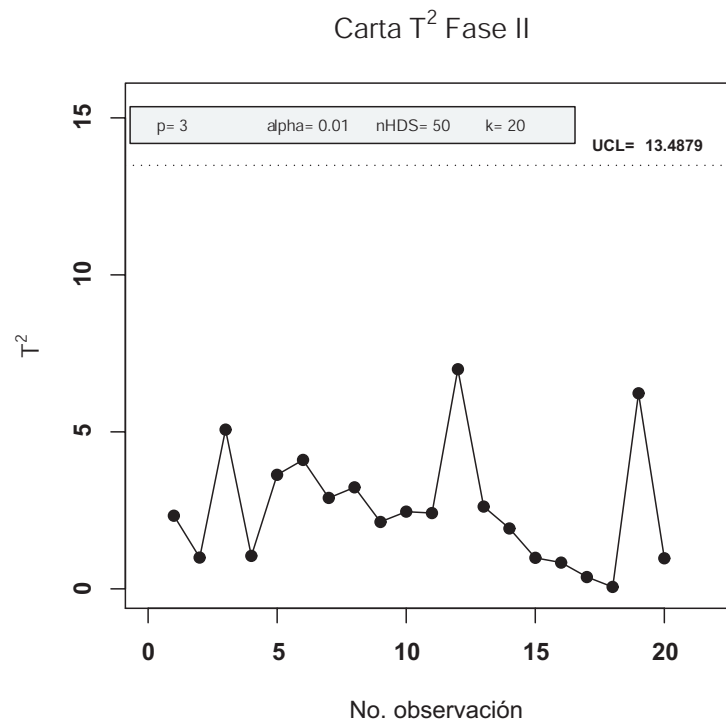


FIGURA 4.14 Carta de control T^2 en la Fase II, con 20 observaciones individuales simuladas. El proceso aparece en control.

Carta de control T^2 para la Fase I con m subgrupos históricos de tamaño n

En este caso se debe usar la ecuación (4.69):

Código R 4.18

```
# Carta T2 con m subgrupos de tamaño n: los datos de los m
# subgrupos son matrices y deben organizarse en un arreglo
de dimensión dim=c(n,p,m)
```

```
carta.T2.subg<-function(X,alpha){
medias<-apply(X, c(3, 2), mean)
var.covar<-apply(X, c(3), var)
m<-dim(X)[3]
p<-dim(X)[2]
n<-dim(X)[1]
media.media<-apply(medias,2,mean)
var.covar.prom<-apply(var.covar,1,mean)
```

```
var.covar.prom<-matrix(var.covar.prom,ncol=p)
T2<-mahalanobis(medias,center=media.media,cov=var.covar.prom)
v1<-p
v2<-m*n-m-p+1
UCL<-(p*(m-1)*(n-1)/(n*v2))*qf((1-alpha),v1,v2)
Subgrupo<-1:m
par(bg="cornsilk")
plot(Subgrupo,T2,type='l',xlim=c(0,m+2),ylim=c(0,max(UCL,
max(T2))+0.5),
main=expression("Carta"*~T^2*~"Fase I"),
ylab=expression(T^2),xlab="No. Subgrupo",font=2)
abline(h=UCL,lty=3)
for(i in 1:m){
temp<-ifelse((T2[i]>UCL),4,19)
points(Subgrupo[i],T2[i],pch=temp)
if(T2[i]>UCL)text(i,T2[i],labels=paste('subg=',i),pos=3,font=2,
cex=0.7)
}
text((max(Subgrupo)-1),UCL,paste('UCL=',round(UCL,digits=4)),
pos=3,font=2,cex=0.7)
legend(locator(1),c(paste("p=",p),paste("m=",m),paste("n=",n),
paste("alpha=",alpha)),ncol=4,cex=0.7,bg='gray95')
(estimaciones<-list(medias=media.media,var=var.covar.prom,
T2=T2))}
```

La función `carta.T2.subg` arroja las estimaciones de los parámetros, los valores T^2 y la gráfica correspondiente. Una aplicación a 40 subgrupos de tamaño 10, de una distribución normal simulada, con $p = 3$, se da a continuación:

Código R 4.19

```
#simulación de los subgrupos:
library(MASS)
mu<-c(28.1,7.18,3.089)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,
0.25),
ncol=3,byrow=T)
a<-matrix(rep(10,40),ncol=40)
X2<-array(apply(a,2,mvrnorm,mu,Sigma),dim=c(10,3,40))

#Aplicación de la función:
res<-carta.T2.subg(X2,0.05)
#NOTA: En la ventana gráfica haga clic donde desee colocar
la leyenda.
```

Salida R 4.15

```
res
$medias
[1] 28.180205 7.415722 3.107282

$var
      [,1]      [,2]      [,3]
[1,] 140.377604 45.015890 1.6097556
[2,] 45.015890 64.011891 3.1948552
[3,] 1.609756 3.194855 0.2199141

$T2
[1] 0.20356322 0.10183617 0.51060289 0.48545652 0.14277642 0.73476718
[7] 0.14561998 0.29399850 0.18889107 0.30089230 0.10938535 0.30965655
[13] 0.38186024 0.54599358 0.20144148 0.02414497 0.40253612 1.16328910
[19] 0.14317715 0.16934141 0.11760159 0.51722483 0.60189684 0.83524049
[25] 0.13083461 0.41509355 0.22577439 0.66050393 0.09275860 0.94643399
[31] 0.45885593 0.27937679 0.20458936 0.17688613 0.21486236 0.08049970
[37] 0.51527826 0.17473582 0.20761257 0.09939644
```

La carta aparece en la figura 4.15.

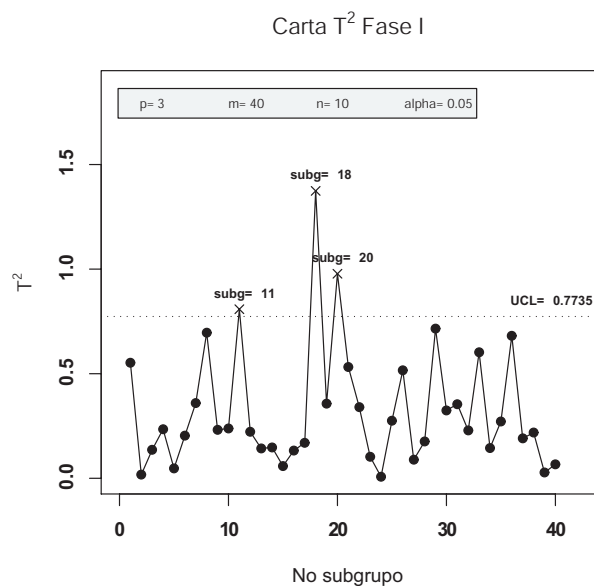


FIGURA 4.15 Carta de control T^2 en la Fase I, con 40 subgrupos simulados de tamaño 10. Los subgrupos 18, 24 y 30 aparecen arriba del UCL, pero no hay causa asignable.

Carta de control T^2 para la Fase II con k subgrupos de tamaño n

La ecuación apropiada para el estadístico T^2 corresponde a (4.68):

Código R 4.20

```
#Carta T2 con subgrupos de tamaño n:
#medias.prom es el vector de medias estimado
#previamente en la Fase I con mHDS subgrupos de tamaño nHDS
#var.covar.prom es la matriz de varianzas covarianzas estimada
#previamente en la Fase I con mHDS subgrupos de tamaño nHDS
#X es el nuevo arreglo de datos recolectados con los cuales
#se verificará el estado del proceso, compuesto por m subgrupos de tamaño n

carta.T2.subgII<-function(X,medias.prom,var.covar.prom,nHDS,mHDS,alpha) {
medias<-apply(X, c(3, 2), mean); p<-dim(X)[2];k<-dim(X)[3]
T2<-mahalanobis(medias,center=medias.prom,cov=var.covar.prom)
v1<-p
v2<-mHDS*nHDS-mHDS-p+1
UCL<-(p*(mHDS+1)*(nHDS-1)/(nHDS*v2))*qf((1-alpha),v1,v2)
Subgrupo<-1:k
par(bg="cornsilk")
plot(Subgrupo,T2,type='l',xlim=c(0,k+2),ylim=c(0,max(UCL,max(T2))+0.5),
main=expression("Carta"~-T^2~- "Fase II"),
ylab=expression(T^2),xlab="No. Subgrupo",font=2)
abline(h=UCL,lty=3)
for(i in 1:k){
temp<-ifelse((T2[i]>UCL),4,19); points(Subgrupo[i],T2[i],pch=temp)
if(T2[i]>UCL)text(i,T2[i],labels=paste('subg=',i),pos=3,font=2,cex=0.7) }
text((max(Subgrupo)-1),UCL,paste('UCL=',round(UCL,digits=4)),pos=3,
font=2,cex=0.7)
legend(locator(1),c(paste("p=",p),paste("m=",mHDS),paste("n=",nHDS),
paste("k=",k),paste("alpha=",alpha)),ncol=5,cex=0.7,bg='gray95')
(resultados<-list(medias=medias.prom,var=var.covar.prom,T2=T2))}
```

La función `carta.T2.subgII` produce los valores T^2 de las nuevas observaciones, y reproduce las estimaciones previas de la matriz de varianzas covarianzas y el vector de medias. Una aplicación: usando las estimaciones obtenidas de la Fase I, se construye la carta T^2 para 20 nuevos subgrupos simulados:

Código R 4.21

```
library(MASS)
Sigma<-matrix(c(140.54,49.68,1.94,49.68,72.25,3.68,1.94,3.68,0.25),
ncol=3,byrow=T)
a2<-matrix(rep(10,20),ncol=20)
X4<-array(apply(a2,2,mvnorm,mu,Sigma),dim=c(10,3,20))

#Construcción carta T2 en la Fase II: se usan los parámetros estimados
#en la Fase I, se suponen guardados en objeto result creado al
#aplicar la función carta.T2.subg

res2<-carta.T2.subgII(X4,res$medias,res$var,10,40,alpha=0.01)
```

Salida R 4.16

```
$medias
[1] 28.180205 7.415722 3.107282

$var
      [,1]      [,2]      [,3]
[1,] 140.377604 45.015890 1.6097556
[2,] 45.015890 64.011891 3.1948552
[3,] 1.609756 3.194855 0.2199141

$T2
[1] 0.18491836 0.21192811 0.04213791 0.54918110 0.48762451 0.69366064
[7] 0.13592005 0.28645523 0.31873985 0.08779386 0.32913567 0.20481764
[13] 0.95397791 0.52623082 0.01122954 0.42978254 0.68318458 0.16662773
[19] 0.08621311 0.31873423
```

La figura 4.16 presenta la carta correspondiente.

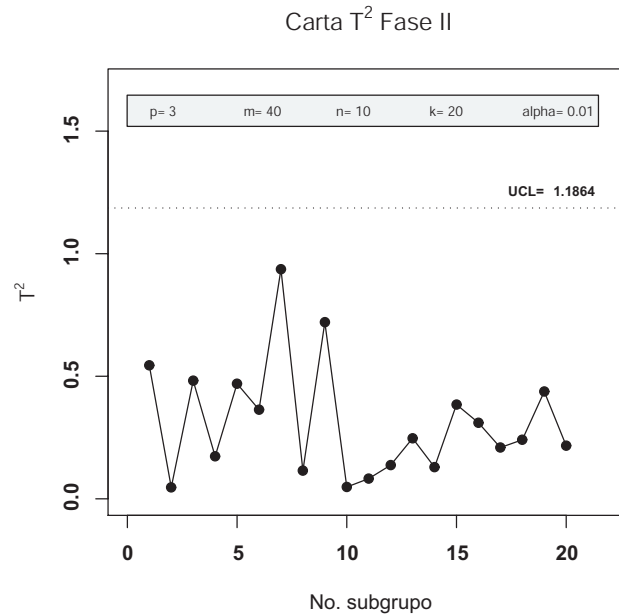


FIGURA 4.16 Carta de control T^2 en la Fase II, con 20 nuevos subgrupos simulados de tamaño 10. El proceso aparece en control.

4.9.2 Construcción de la base de datos históricos o HDS

El conjunto de datos históricos o de línea base se emplea en la Fase I o fase retrospectiva de análisis del proceso multivariado. El propósito fundamental es determinar si en ese pasado el proceso estuvo estable, y obtener estimaciones preliminares de los parámetros y del límite de control para la carta T^2 o UCL.

En la Fase I se realiza una depuración de este conjunto de datos históricos mediante la identificación de observaciones que no se acomodan a la estructura general o definida por la mayoría de los datos. Estas observaciones son etiquetadas como posibles *outliers*. Cuando hay evidencia empírica de que son realmente *outliers*, son removidos del conjunto de datos históricos, se procede a reestimar los parámetros del proceso, y este procedimiento se repite cuantas veces sea necesario, es decir, hasta que el conjunto final de observaciones no contenga observaciones atípicas. Al

conjunto de datos finalmente obtenido se le denomina HDS. Obviamente las condiciones de tamaño de muestra citadas en la sección 4.8.3 son aplicables al HDS.

Una suposición que es importante verificar es la independencia de las observaciones, pues ignorarla puede afectar seriamente las estimaciones de las varianzas y covarianzas del proceso, así como el UCL. Cuando tal situación está presente se deben realizar ajustes apropiados en las estimaciones. Las fases en la construcción del HDS son las siguientes:

Planeación del control

Entre otras cosas, en esta etapa se deben establecer las variables por controlar, básicamente identificando dónde existen problemas o dónde estos pueden aparecer con serias consecuencias. Para ello es necesario conocer el proceso, lo cual se puede lograr construyendo un diagrama del mismo, y luego de analizarlo en detalle, seleccionando las variables de control; así mismo se deben definir en cada caso las condiciones de operación que se consideran “buenas” y los propósitos del control que se desea implementar en el proceso en general, al igual que en cada una de las variables involucradas: mantener el proceso tan cerca como sea posible de un valor objetivo, minimizar la variabilidad, etc.

También es importante definir la frecuencia, la forma del muestreo y las unidades de medida para cada variable.

Datos preliminares

Los datos del HDS deben proporcionar una imagen lo más precisa posible del estado estable del proceso. Una vez se tiene recolectada la base de datos bajo las condiciones de planeación adecuadas, se procede a realizar los estudios estadísticos descriptivos univariados (histogramas, medidas de resumen, gráficos de series de tiempo de cada variable, etc.) y multivariados (gráficos y matrices de dispersión, medidas de correlación, etc.).

También se deben llevar a cabo verificaciones sobre la independencia entre observaciones. Una descripción práctica de este tipo de verificación se puede ver en Mason & Young (2002, pp. 57-61).

4.9.3 Procedimientos de recolección de datos

El procedimiento para la recolección de datos debe tratar de garantizar, simultáneamente, el mínimo error de registro y datos en tiempo real de las variables observadas, sin que la frecuencia de registro sea tan corta que impida observar variación, ni tan larga que se pierda información sobre acontecimientos críticos en el proceso.

4.9.4 Datos faltantes

La manera más simple de manejar datos faltantes es eliminando el registro completo, pero tal procedimiento solo resulta apropiado si la muestra o conjunto final sigue siendo representativo del proceso. A nivel multivariado se han planteado varios procedimientos para imputar (remplazar) el o los datos faltantes. Uno de tales procedimientos está sustentado en la predicción del dato faltante con base en la regresión de la variable en la cual se sitúa dicho dato *vs.* el resto de variables del proceso. Por ejemplo, considere los datos simulados de la tabla 4.3 que corresponden a medidas tomadas sobre seis dimensiones expresadas en mm, de 50 pasadores de aluminio que se usan en aviones, listados según orden de producción, donde:

- X1: diámetro 1
- X2: diámetro 2
- X3: diámetro 3
- X4: diámetro de cabeza
- X5: longitud sin cabeza
- X6: longitud con cabeza

TABLA 4.3 Medidas simuladas de 50 pasadores de aluminio usados en aviones.

Pasador	X1	X2	X3	X4	X5	X6
1	10,339	10,376	10,250	15,599	47,314	57,315
2	9,632	9,771	9,758	14,688	51,059	58,924
3	10,068	10,047	9,929	15,174	49,188	59,144
4	10,103	9,877	9,724	15,089	48,154	58,944
5	10,381	10,374	10,787	15,260	48,916	62,583
6	10,321	10,284	10,235	15,481	48,758	58,578
7	9,473	9,431	9,434	14,527	48,939	58,618
8	10,818	10,588	10,770	15,654	50,816	61,049
9	10,323	10,418	10,202	15,276	48,119	60,254
10	10,179	10,184	10,048	15,095	50,670	60,757
11	10,393	10,252	10,385	15,540	49,292	59,892
12	10,217	10,265	10,236	14,927	48,677	58,916
13	10,660	10,756	10,586	15,411	49,975	61,203
14	9,862	9,963	9,703	14,682	49,113	60,279
15	9,087	9,151	8,976	13,893	50,560	60,777
16	9,197	9,278	9,485	14,252	48,684	60,739
17	9,458	9,571	9,363	14,179	48,394	59,271
18	9,576	9,443	9,653	14,003	48,877	60,790
19	10,004	9,993	10,199	15,241	49,830	59,618
20	10,269	10,164	10,210	15,524	50,123	59,595
21	9,431	9,264	9,045	14,324	50,997	61,089
22	9,686	9,751	9,831	15,414	50,928	60,653
23	10,195	10,311	10,369	15,435	49,251	60,225
24	9,871	10,109	10,168	14,988	49,360	59,341
25	9,934	10,024	10,021	14,915	51,540	62,261
26	9,621	9,555	9,311	14,578	47,776	57,579
27	9,860	10,248	10,078	14,751	50,206	60,503
28	10,554	10,720	10,611	15,468	48,740	60,193
29	11,299	11,094	10,898	16,251	48,120	59,356
30	10,431	10,658	10,519	15,365	47,025	58,125
31	8,966	8,959	8,893	13,867	50,229	59,133
32	9,746	9,880	9,911	15,127	49,466	58,694
33	10,268	10,519	10,354	15,341	50,416	60,067
34	9,120	9,257	9,069	13,931	49,786	59,044
35	9,802	10,067	9,922	14,800	46,414	57,612
36	9,960	10,164	9,912	14,948	48,742	60,117
37	9,794	9,859	9,869	14,976	49,597	59,966
38	9,287	9,261	9,286	14,431	49,465	58,743
39	9,622	9,613	9,814	14,405	48,946	59,283
40	10,137	9,909	9,961	14,985	48,246	58,400
41	9,425	9,420	9,281	14,198	49,310	58,640
42	10,248	9,871	10,239	15,115	51,015	60,169
43	10,295	10,311	10,041	15,025	49,253	60,748
44	10,727	10,549	10,668	15,806	48,054	58,429
45	10,303	10,566	10,427	15,399	49,650	60,970

TABLA 4.3 (continuación)

46	10,061	10,186	10,302	15,005	49,496	59,479
47	10,145	10,282	10,137	15,231	48,114	59,268
48	9,585	9,526	9,453	14,437	50,147	61,130
49	10,778	10,708	10,917	15,930	49,188	60,883
50	9,800	9,783	10,000	15,101	51,131	61,511

Suponga que faltan los datos de la variable X3 en el registro 20 y de la variable X5 en el registro 30. Se eliminan inicialmente las filas 20 y 30 de la base de datos, y se hacen las siguientes regresiones con los 48 datos restantes: X3 *vs.* X1, X2, X4, X5, y X6 y X5 *vs.* X1, X2, X3, X4, X6:

Código R 4.22

```
#Cargando objeto R donde está la base de datos
#source(file.choose()) permite navegar en disco local hasta
ubicar archivo
#datospasadores.txt, tal archivo contiene el objeto R
pasadores, source(file.choose())
```

```
# Eliminando filas 20 y 30
datos1<-pasadores[c(-20,-30),]
# Regresión de X3 vs. X1, X2, X4, X5, y X6:
regres1<-lm(X3~X1+X2+X4+X5+X6,data=datos1)
summary(regres1); nf<-layout(rbind(c(1,1,2,2),c(3,3,4,4)));
plot(regres1)
```

Salida R 4.17

```
Call: lm(formula = X3 ~ X1 + X2 + X4 + X5 + X6, data = datos1)
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.331420 -0.097263 -0.008809  0.092368  0.314061
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.60192    1.40002  -2.573  0.01371 *
X1           0.23735    0.17283   1.373  0.17693
X2           0.48276    0.15682   3.078  0.00366 **
X4           0.25196    0.11338   2.222  0.03170 *
X5          -0.01031    0.02905  -0.355  0.72435
X6           0.05213    0.02688   1.939  0.05918 .
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.153 on 42 degrees of freedom
Multiple R-Squared: 0.9203,    Adjusted R-squared: 0.9108
F-statistic: 96.95 on 5 and 42 DF,  p-value: < 2.2e-16
```

La ecuación ajustada según la salida R 4.17 es:

$$\hat{X}_3 = -3,60192 + 0,23735X_1 + 0,48276X_2 + 0,25196X_4 - 0,01031X_5 + 0,05213X_6,$$

con la cual se procede a predecir el valor de X_3 en la fila 20, con las demás variables tomando los valores de estas en dicha fila:

Salida R 4.18

pasadores[20,] #Valores en los datos completos, registro 20

	X1	X2	X3	X4	X5	X6
20	10.269	10.164	10.21	15.524	50.123	59.595

Con la ecuación ajustada, el valor estimado para X_3 en la observación 20 es de 10,24355, que comparado con el verdadero valor, 10,210, da una muy buena estimación. Para X_5 , se procede con la regresión así:

Código R 4.23

```
regres2<-lm(X5~X1+X2+X3+X4+X6,data=datos1)
summary(regres2)
nf<-layout(rbind(c(1,1,2,2),c(3,3,4,4)))
plot(regres2)
```

Salida R 4.19

Call: lm(formula = X5 ~ X1 + X2 + X3 + X4 + X6, data = datos1)

Residuals:

	Min	1Q	Median	3Q	Max
	-1.79658	-0.40725	-0.05406	0.48102	2.09990

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.7681	7.7805	1.513	0.1379
X1	-0.7177	0.9305	-0.771	0.4448
X2	-0.9812	0.9083	-1.080	0.2862
X3	-0.2901	0.8172	-0.355	0.7243
X4	1.2321	0.6067	2.031	0.0486 *
X6	0.6521	0.1096	5.948	4.71e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8113 on 42 degrees of freedom

Multiple R-Squared: 0.5099, Adjusted R-squared: 0.4516

F-statistic: 8.74 on 5 and 42 DF, p-value: 9.602e-06

pasadores3[30,]

	X1	X2	X3	X4	X5	X6
30	10.431	10.658	10.519	15.365	47.025	58.125

La ecuación ajustada es:

$$\hat{X}_5 = 11,7681 - 0,7177X_1 - 0,9812X_2 - 0,2901X_3 + 1,2321X_4 + 0,6521X_6$$

y así se puede predecir el valor de X_5 en la fila 30, con los valores de las demás variables en dicha fila. Por tanto, el valor predicho para X_5 , con las observaciones de las demás variables en la fila 30 es 47,60711. Compare con el verdadero valor de 47,025.

En las figuras 4.17 y 4.18 se presentan, respectivamente, los gráficos de los residuales para las dos anteriores regresiones. Como se puede ver, hay un buen comportamiento de estos en ambos casos. El anterior método, sin embargo, depende de qué tan buena sea la ecuación de predicción ajustada.

Otros métodos han sido desarrollados para la imputación de datos faltantes, tales como el algoritmo EM. Una breve descripción de este algoritmo se da a continuación.

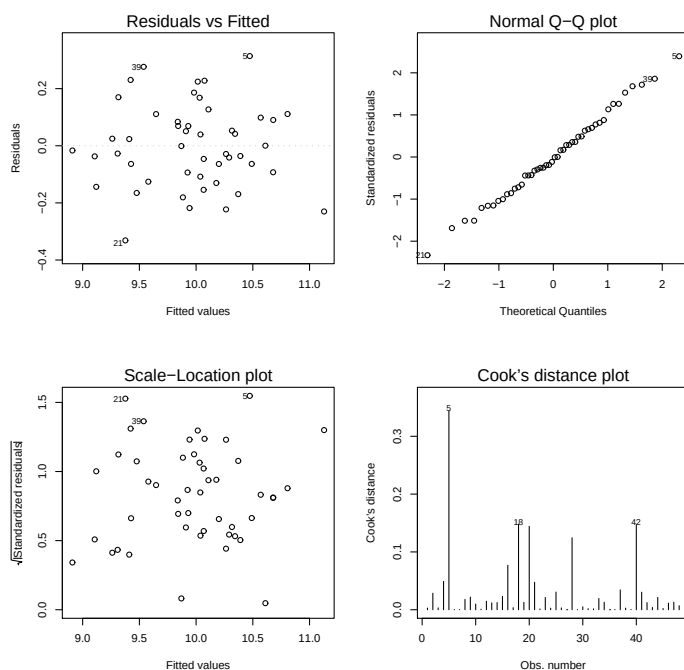


FIGURA 4.17 Gráficos de residuales para la regresión de X_3 vs. X_1, X_2, X_4, X_5 y X_6 .

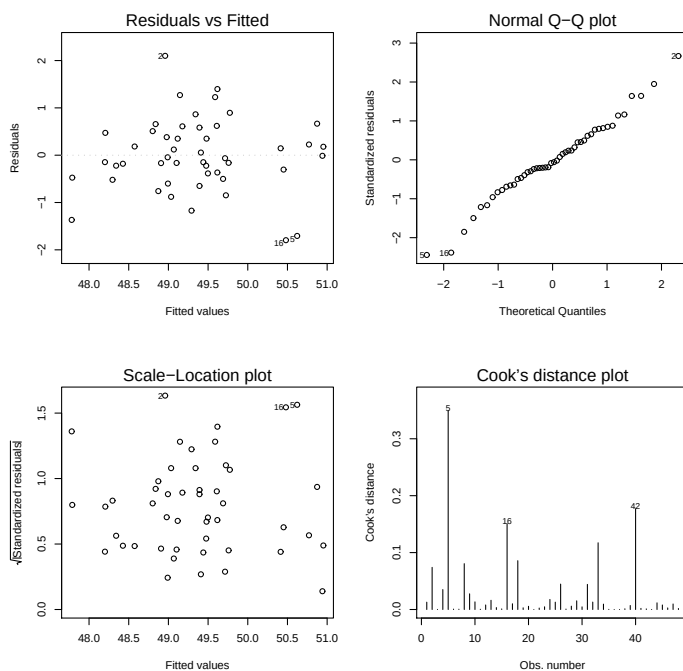


FIGURA 4.18 Gráficos de residuales para la regresión de X_5 vs. X_1 , X_2 , X_3 , X_4 y X_6 .

El algoritmo EM (*Expectation - Maximization Algorithm*) es un algoritmo general para estimación por máxima verosimilitud donde los datos son incompletos o la función de verosimilitud incluye variables latentes. El algoritmo inicia asignando valores de acuerdo con algún conocimiento a priori sobre los valores óptimos, a todos los parámetros por ser estimados; luego, se mueve iterativamente entre dos pasos, el primero llamado “E-Step” (*Expectation Step*) y el segundo llamado “M-Step” (*Maximization Step*).

En el paso “E-Step”, el algoritmo calcula la verosimilitud esperada de los datos completos, o función Q, y estima los estadísticos suficientes para los datos completos dados los datos observados y los valores actuales de los parámetros (es decir, la esperanza es tomada sobre la distribución condicional de las variables latentes dados los valores actuales ajustados de los parámetros y los datos observados). Cuando el log de verosimilitud con los datos completos es una función lineal de las variables no observadas, entonces lo anterior es equivalente a hallar el valor esperado de los

datos no observados dados los datos observados y los valores actuales de los parámetros.

En el paso “M-Step” se maximiza la verosimilitud asociada con los estadísticos estimados en el paso previo, es decir, reestima todos los parámetros maximizando la función Q , usando los datos estimados. Una vez que se tiene un nuevo conjunto de valores de los parámetros, se repiten los pasos “E-Step” y “M-Step” hasta que la verosimilitud alcance un máximo local (es decir, converge), o equivalentemente cuando $\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\| < \varepsilon$, donde $\boldsymbol{\theta}^t$ es el vector de parámetros estimado en la iteración t .

Viendo lo anterior de una manera sencilla, el algoritmo M procede iterativamente aumentando los datos mediante valores supuestos de las variables ocultas o latentes y reestima los parámetros asumiendo que los valores supuestos son los verdaderos.

En el algoritmo EM solo puede garantizarse la convergencia a un máximo local y no global (los estimadores de máxima verosimilitud). Si existen varios máximos locales, la convergencia al máximo global depende de los valores iniciales. Por lo anterior, algunas estrategias recomiendan probar con muchos valores iniciales y elegir la solución que converge al valor de verosimilitud más alto; otras usan un modelo más simple para determinar un valor inicial para modelos más complejos. Para detalles técnicos, ver Johnson & Wichern (1998).

4.9.5 Detección de colinealidad

Esta situación puede ocurrir a causa de una fuerte asociación entre dos o más de las variables consideradas en el monitoreo, es decir, por multicolinealidad. Estas asociaciones pueden ser debidas a relaciones de causa-efecto, presencia de *outliers* y restricciones del muestreo (en muestras pequeñas puede haber más posibilidad de singularidad que en muestras grandes).

¿Cómo diagnosticar la colinealidad? Analizando la matriz de correlaciones, evaluando los factores de inflación de varianza o VIF (que mide el efecto combinado de la dependencia entre las variables sobre la varianza de la variable para la cual se calcula), analizando los valores y

vectores propios de la matriz de varianzas covarianzas (valores propios pequeños implican dependencia aproximadamente lineal en las columnas de la matriz de diseño) o estableciendo el número de condición de la matriz de varianzas covarianzas. Para más detalle de estos análisis, ver, por ejemplo, Montgomery & Peck (1992).

Entre las soluciones a este problema, en el ámbito del control multivariado, se ha aplicado la técnica PCA o de análisis de componentes principales, la cual, además de ayudar a detectar la existencia de singularidad, también permite reducir la dimensionalidad reagrupando las variables que están altamente correlacionadas (ver Apéndice para más detalles sobre el análisis de componentes principales). Otra estrategia consiste en eliminar alguna de las variables involucradas en una multicolinealidad severa. Algunas reglas prácticas para determinar cuál variable eliminar son: comenzar por aquellas variables que sean obtenidas mediante combinación de información de otras variables del proceso, y remover aquellas que estén asociadas a valores propios casi nulos de la matriz de varianzas covarianzas. Estas aproximaciones se deben usar con cuidado, puesto que al reducir el número de componentes principales se puede perder la habilidad para identificar desviaciones en algunas direcciones.

4.9.6 Diagnóstico de no independencia entre observaciones

Cabe recordar que un supuesto fundamental en el control del proceso es que las observaciones sucesivas sobre las variables son independientes; sin embargo, este supuesto puede no cumplirse por varias razones, las cuales se pueden clasificar en:

1. Debidas a relaciones de causa-efecto: los valores actuales de la variable son proporcionales a valores pasados de la misma, es decir, determinados por dichos valores.
2. Debidas a asociación: los valores actuales de la variable solo se asocian a valores pasados, sin que estos últimos determinen los primeros, y en este caso generalmente hay una variable latente o no observada con la cual se halla asociada la variable que se examina, de modo que, al ignorar tal situación, se hace aparente la existencia de tendencias en las observaciones de las variables medidas con las cuales dicha variable latente se halla asociada.

Ahora bien, la forma en que ocurre la correlación también puede clasificarse en dos:

1. Como un decaimiento continuo u uniforme, cuando la observación actual depende de algún valor pasado cercano; por ejemplo, en proceso de filtrado en línea de impurezas, medidas sobre el coeficiente de transferencia de calor.
2. Decaimiento de escalón, que ocurre cuando el cambio en el tiempo de la variable es inconsistente sobre periodos cortos, pero en periodos más extendidos se observa un cambio de escalón puede ocurrir en procesos donde el cambio con el tiempo ocurre muy lentamente. En este caso, el desempeño en una etapa está en dependencia con el desempeño en etapas previas.

Las consecuencias de la autocorrelación se traducen en observaciones distorsionadas del desempeño de los procesos, y a nivel multivariado las soluciones llegan a ser más complicadas que las hasta ahora propuestas a nivel univariado, dado que no solo se debe atender a las relaciones de las observaciones en el tiempo sobre una misma variable, sino también a cómo las variables se relacionan con otras variables del proceso en el tiempo. La autocorrelación no elimina estas relaciones pero puede confundirlas, y por tanto debe ser removida para lograr una interpretación apropiada de las señales. El tratamiento de esta situación se discute en detalle en Mason & Young (2002).

A continuación se describirán algunos métodos para chequear autocorrelación. A nivel univariado, con cada variable del proceso, se puede verificar mediante el gráfico de valores actuales *vs.* valores rezagados k lugares en el tiempo, la función de autocorrelación o ACF, la función de autocorrelación parcial o PACF y la prueba Lung-Box. En la tabla 4.4 se presentan los valores de 300 observaciones de la serie de tiempo (datos por hora) de una variable z_t simulada como un proceso AR(1). Se pueden obtener gráficas para analizar estos datos univariados mediante las siguientes rutinas en R:

Código R 4.24

```
# Para generar un proceso AR(1), se puede utilizar el siguiente
programa
res<-NA; z<-NA; n<-300; zt.1<-17 #valor inicial (t-1)
phi<-0.4
```

```

for(t in 1:n) {
  z[t]<-phi*zt.1+rnorm(1,mean=0,sd=1)
  zt.1<-z[t]
  res<-append(res,zt.1) }
res<-res[-1]
zt<-res
# Con este zt puede correr el siguiente programa
# Se entran los 300 datos en orden de tiempo, en un objeto
vectorial
zt<-scan()
15.78 15.14 16.44 18.22 18.11 18.78 17.26 14.79 16.37 16.28
15.7 15.16 15.92 16.02 17.51 15.99 17.13 16.19 17.62 17.79
16.32 17.27 17.35 18.09 18.39 17.86 15.95 16.32 15.43 16.61
15.54 16.46 18.87 16.92 17.48 15.96 16.27 17.87 17.44 17.34
.
.
14.41 14.79 16.65 16.37 15.82 17.15 16.78 18.75 18.37 18.89
17.15 17.81 17.15 16.29 16.81 17.78 16.72 16.48 16.84 16.34
17.11 18.38 17.15 16.97 15.07 17.02 16.09 16.85 16.03 16.36
16.5 16.45 15.73 17.02 15.91 15.3 16.52 16.99 18.32 18.93

# Los datos ingresados son convertidos a un objeto 'ts'.
La serie inicia
# en t=1 y la frecuencia es 1:

zt<-ts(zt,start=1,frequency=1)

win.graph()
# Dividir ventana gráfica en tres secciones de igual ancho
y alto
nf<-layout(rbind(c(0,1,1,0),c(2,2,3,3)))

# Graficar serie
plot(zt, main=expression(paste("Proceso AR(1) simulado",
sep=" ",phi==0.4))
,xlab="Tiempo")

# Graficar ACF para k=1,2,...,36
q<-acf(zt, lag.max=36,plot=TRUE,type="correlation",xaxt="n",
main=expression(paste("ACF proceso AR(1) simulado",sep=" ",
phi==0.4)))
axis(1,at=q$lag, labels=as.character(c(0:36)),cex.axis=0.7)

# Grafica PACF k=1,2,...,36
p<-pacf(zt, lag.max=36,plot=TRUE,xaxt="n",main=expression
(paste("PACF proceso AR(1) simulado",sep=" ",phi==0.4)))
axis(1,at=q$lag, labels=as.character(c(0:36)),cex.axis=0.7)
# Graficar serie vs. rezagos k=1,2,..6
win.graph()

```

```
lag.plot(zt, lags=6, diag=TRUE, diag.col="blue", do.lines=FALSE,
main=expression(paste("Proceso AR(1) simulado", sep=" ",
phi==0.4)), cex.main=0.8 )
```

Las gráficas correspondientes a la serie de tiempo, la ACF y la PACF, aparecen en la figura 4.19, y en la figura 4.20, la gráfica de rezagos. Una función para obtener los resultados de la prueba Lung-Box aparece a continuación:

Código R 4.25

```
lung.box.test<-function(series,maxlag){
aux<-floor(maxlag/6)
X.squared<-c(rep(NA,aux))
df<-c(rep(NA,aux))
p.value<-c(rep(NA,aux))
for(i in 1:aux){
lung<-Box.test(zt, lag=(6*i), type="Ljung-Box")
X.squared[i]<-lung[[1]]
df[i]<-lung[[2]]
p.value[i]<-lung[[3]] }
lag<-6*c(1:aux)
lungbox<-as.data.frame(cbind(X.squared,df,p.value))
rownames(lungbox)<-lag; lungbox }
```

TABLA 4.4 Datos serie simulada. Orden de observaciones:
por fila.

15,78	15,14	16,44	18,22	18,11	18,78	17,26	14,79	16,37	16,28
15,70	15,16	15,92	16,02	17,51	15,99	17,13	16,19	17,62	17,79
16,32	17,27	17,35	18,09	18,39	17,86	15,95	16,32	15,43	16,61
15,54	16,46	18,87	16,92	17,48	15,96	16,27	17,87	17,44	17,34
17,46	17,43	18,91	17,39	15,54	14,24	15,58	15,72	15,56	13,83
14,93	16,04	17,87	17,83	17,07	17,57	17,44	17,24	17,54	17,78
16,85	16,49	17,30	16,86	17,30	17,59	15,44	16,45	17,39	17,11
15,76	15,92	14,79	15,53	15,09	16,15	16,25	14,82	14,58	16,01
17,05	15,92	17,32	18,78	18,41	18,60	17,35	18,04	16,98	16,35
15,21	15,21	16,26	16,15	18,01	17,68	15,89	17,23	17,19	17,42
16,30	18,80	18,17	16,98	17,42	17,41	17,63	20,02	16,98	16,61
15,67	16,03	14,30	17,48	16,68	15,85	17,17	19,31	16,81	15,72
16,62	17,67	15,97	15,84	16,27	16,26	16,06	16,97	16,31	16,50
16,24	16,12	17,64	16,49	16,66	16,44	15,61	16,37	17,88	17,02
17,72	16,81	16,67	14,29	17,20	17,30	17,47	18,61	17,05	16,63
16,53	16,15	16,92	17,43	17,96	16,53	16,41	17,91	17,63	17,28
15,85	16,04	16,22	18,00	18,55	18,87	16,81	17,17	17,43	15,68
15,85	16,62	15,37	14,86	16,80	16,90	16,63	17,09	17,29	16,95
16,60	16,71	15,66	15,25	14,64	16,57	15,12	15,90	16,13	17,02
15,82	15,89	14,79	16,04	17,79	18,25	16,78	15,44	16,55	17,03

TABLA 4.4 (continuación)

18,45	17,71	15,70	16,83	17,20	16,36	17,09	16,38	16,65	15,41
16,80	15,61	18,04	16,99	16,45	14,40	15,71	16,11	15,82	15,43
15,78	15,28	16,14	17,19	15,05	15,90	15,63	16,05	16,72	17,19
17,48	16,56	16,57	15,82	15,71	17,19	17,97	19,07	17,02	17,64
16,99	17,34	16,89	15,74	16,45	16,15	16,55	17,70	17,17	19,03
18,48	18,78	17,41	15,59	5,26	16,38	16,72	15,84	16,95	14,98
14,41	14,79	16,65	16,37	15,82	17,15	16,78	18,75	18,37	18,89
17,15	17,81	17,15	16,29	16,81	17,78	16,72	16,48	16,84	16,34
17,11	18,38	17,15	16,97	15,07	17,02	16,09	16,85	16,03	16,36
16,50	16,45	15,73	17,02	15,91	15,30	16,52	16,99	18,32	18,93

Salida R 4.20

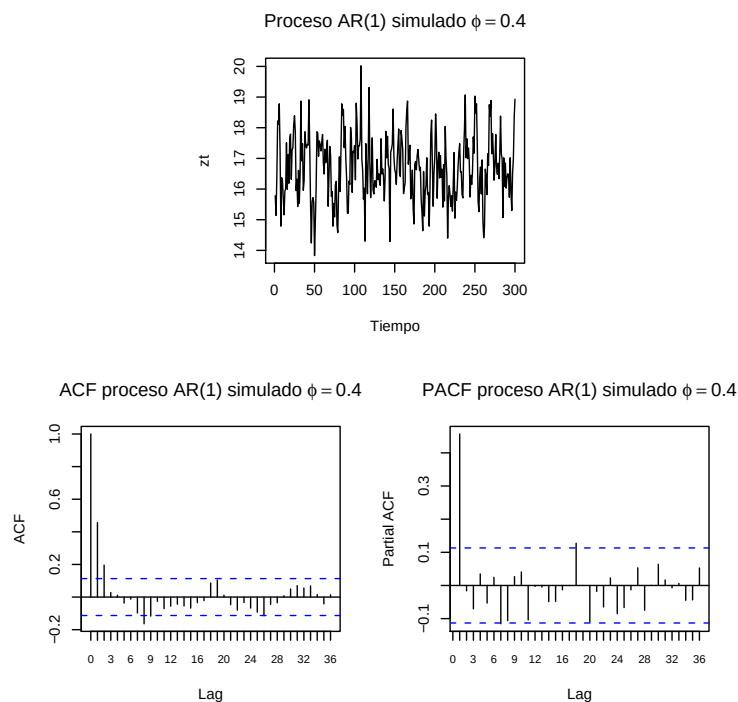


FIGURA 4.19 Serie simulada, su ACF y su PACF. La serie presenta una autocorrelación significativa, aunque es estacionaria y, según la ACF y la PACF, puede modelarse como un AR(1).

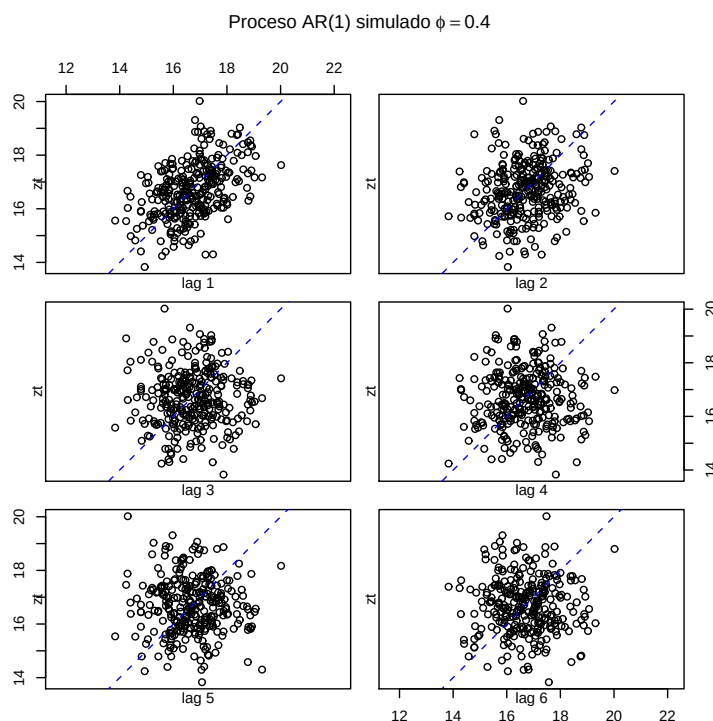


FIGURA 4.20 Gráfico de la serie z_t vs. sus rezagos, $k = 1, 2, \dots, 6$. Observe que entre z_t y su rezago de orden 1 existe una clara asociación lineal, pero en rezagos superiores dicha asociación es débil o prácticamente inexistente. Por tanto, la variable puede ser modelada como un proceso AR(1).

Aplicando la prueba Lung-Box a los datos de la serie, se obtiene para $k = 6, 12, 18, 24$ y 36 el rechazo de la hipótesis de ruido blanco, es decir, hay autocorrelación:

Salida R 4.21

```
lung.box.test(zt, 36)
      X.squared df      p.value
6    75.68871  6 2.764455e-14
12   93.98853 12 8.326673e-15
18   99.90676 18 2.303713e-13
24  108.08791 24 1.214695e-12
30  117.10124 30 3.108958e-12
36  122.12152 36 2.671008e-11
```

(Se recomienda al lector no familiarizado con el análisis de series de tiempo ver el texto de Guerrero, 2003).

Sin embargo, los anteriores análisis pueden ser numerosos si el proceso tiene una dimensión muy grande. Una recomendación antes de proceder de la anterior forma consiste en agregar la variable de secuencia de tiempo en la matriz de diseño, calcular la matriz de correlación sobre esta matriz ampliada y evaluar las correlaciones de las variables con la variable de secuencia de tiempo. Obviamente, correlaciones significativas con esta última variable son claras indicaciones de que existe dependencia entre las observaciones de la(s) variable(s) que presenta tal asociación con el tiempo. Sin embargo, correlaciones no significativas con el tiempo no indican que no haya autocorrelación entre las observaciones de las variables. Por tanto, para aquellas variables para las cuales no se detecta asociación lineal con la variable de secuencia de tiempo, se recomienda realizar el análisis de la ACF, PACF y Lung-Box, pues los procesos AR, por ejemplo, son estacionarios en covarianza, por tanto su media es constante y no presentarían asociación lineal significativa con la variable de secuencia de tiempo. Para la serie de datos simulados, previamente presentados en esta sección, se calcula la correlación con la variable $t = 1, 2, \dots, 300$. Los resultados obtenidos con el paquete R se dan a continuación:

Código R 4.26

```
t<-1:300
cor.test(zt,t,method="pearson",conf.level = 0.95)
```

Salida R 4.22

```
Pearson's product-moment correlation
data:  zt and t
t = -0.1709, df = 298, p-value = 0.8645

alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1229996  0.1034602

sample estimates:
      cor
-0.009896625
```

Es evidente que no existe asociación lineal significativa entre z_t y t ; no obstante, el proceso presenta autocorrelación de orden 1, como fue demostrado previamente.

4.10 Fase I de control

Esta fase, también conocida como fase retrospectiva, tiene como objetivo principal la depuración de la base de datos histórica, para establecer el HDS o conjunto de datos históricos de base por usar en la estimación de los parámetros del proceso multivariado. Básicamente, trata de identificar observaciones en las cuales se produjeron situaciones de fuera de control y eliminar los posibles *outliers*; a nivel univariado, los *outliers* suelen presentarse como observaciones extremas o alejadas de la gran mayoría de los datos; sin embargo, a nivel multivariado se hace más difícil tratar de caracterizar un *outlier* en función de su lejanía, dado que una observación en un espacio p dimensional puede ser extrema en cierta dirección pero no en otras direcciones; además, deben considerarse como observaciones alejadas del conjunto principal de datos, no solo por la magnitud de la distancia estadística al centro de dicha masa principal, sino también porque pueden seguir un patrón diferente al de la mayoría de las observaciones debido a que “son miembros de una población diferente del modelo principal introducidos dentro de la muestra” (Rousseeuw & Leroy, 1987).

Algunos de los métodos multivariados para detección de *outliers* están orientados a realizar proyecciones de los datos en diferentes direcciones, y tratan de identificar si alguna de estas direcciones muestra a una observación como *outlier*, en tanto que los métodos multivariados de estimación robusta construyen estimaciones asignando pesos a las observaciones en proporción inversa a sus “lejanías” de la masa principal.

El estadístico T^2 en sí mismo no es un método óptimo para la identificación de *outliers*, particularmente debido a que está basado en dos estadísticos muestrales ($\bar{\mathbf{x}}$ y $\mathbf{S}_u$ o $\mathbf{S}_n$) que sufren de un fenómeno conocido como enmascaramiento: inhabilidad de un procedimiento para señalar incluso a un único *outlier* en presencia de múltiples *outliers* multivariados, ubicados arbitrariamente dentro del conjunto de datos.

¿Qué problemas se pueden derivar de la no identificación de *outliers*? Básicamente, las estimaciones resultantes son sesgadas, el riesgo de no solucionar este problema es establecer estimaciones y un UCL que no den cuenta del verdadero rango de variación inherente del proceso, y posiblemente afectar desfavorablemente la potencia del procedimiento para la detección de otros tipos de desviaciones en el proceso, así como

la tasa de falsas alarmas generadas. Gnanadesikan (1977) establece que los *outliers* multivariados pueden resultar críticos debido a sus complejas consecuencias: “Un *outlier* multivariado puede distorsionar no solo las medidas de localización y escala, sino también las de orientación (la correlación)”. En consecuencia, los resultados de pruebas estadísticas y de las estimaciones resultantes de los parámetros de interés pueden llegar a ser inestables, teniendo en cuenta que la cantidad y magnitud de este tipo de observaciones varían de muestra a muestra; por esta razón resulta necesario que los procedimientos de estimación multivariada presenten robustez ante la presencia de múltiples *outliers* multivariados.

Para el caso particular de construcción de una carta robusta en la Fase I monitoreando sobre n observaciones individuales, ver Yáñez *et al.* (2003) y Vargas (2003).

El proceso de depuración que se lleva a cabo en la Fase I inicia fijando la tasa de falsas alarmas o probabilidad de error tipo I para la carta en la Fase I. Si esta tasa es demasiado grande, se puede llegar a catalogar como *outliers* a observaciones buenas, lo cual tiene consecuencias importantes sobre la calidad de las estimaciones cuando la base de datos es pequeña. Por el contrario, si esta tasa es muy pequeña, se puede estar incluyendo observaciones *outliers* y de nuevo la calidad de las estimaciones se verá afectada en muestras pequeñas.

4.10.1 Depuración bajo normalidad

Sea en el caso de monitoreo con observaciones individuales o sobre subgrupos racionales, este proceso consta del siguiente esquema iterativo:

1. En este primer paso, estimar el vector de medias, la matriz de varianzas covarianzas y el UCL para la carta, de acuerdo con la distribución apropiada para el estadístico T^2 .
2. Calcular los valores T_i^2 y compararlos con el UCL.
3. Si hay valores $T_i^2 > UCL$, investigar para determinar si existen causas asignables. Eliminar de la base de datos aquellas observaciones con causas asignables removibles. Conservar aquellas observaciones señalantes pero sin causas asignables.

4. Con el conjunto de datos que se conservan, reestimar los parámetros y el UCL.
5. Repetir 2 a 4 hasta no hallar observaciones por fuera de límites con causas asignables. Es posible que al concluir el proceso existan observaciones tales que $T_i^2 > UCL$, pero sin causas asignables; sin embargo, no se remueven del conjunto por ser parte de la distribución del proceso en control, y en las cuales se ha incurrido en falsa alarma.

Mason & Young (2002) recomiendan, en el caso de monitoreo sobre subgrupos racionales, ignorar la organización en tales subgrupos y considerar que se tiene un solo conjunto de datos de tamaño nm y proceder con ellos como en el caso del monitoreo sobre observaciones individuales, para depurar la base de datos de la presencia de *outliers*. Si el proceso está en control, esta aproximación produce el estimador más eficiente de la matriz de varianzas covarianzas.

Otro método para identificar *outliers* se basa en el uso del gráfico Q-Q apropiado a la distribución del estadístico T^2 en la Fase I. Aquellos puntos que no caen sobre la línea recta son candidatos a *outliers*.

4.10.2 Depuración bajo no normalidad

Bajo no normalidad, previamente se describieron algunos métodos no paramétricos para la estimación del UCL de la carta T^2 . A continuación se detalla el procedimiento de depuración de la base de datos con cada uno de dichos procedimientos.

Aproximación de Chebyshev

1. En este primer paso, estimar el vector de medias y la matriz de varianzas covarianzas.
2. Calcular los valores T_i^2 para las observaciones del conjunto de datos históricos, y obtener su media ($\bar{T}$) y su desviación estándar (S_T).
3. Aproximar el UCL como $\bar{T} + kS_T$, con $k = \sqrt{\alpha^{-1}}$, de acuerdo con la ecuación (4.70).

4. Comparar los valores T_i^2 con el anterior UCL estimado. Para aquellos valores para los cuales $T_i^2 > UCL$, determinar si existen causas asignables removibles, y eliminar las observaciones correspondientes del conjunto de datos.
5. Con las observaciones que queden, recalcular los estimadores muestrales del vector de medias y la matriz de varianzas covarianzas. Repetir 2 a 4 hasta obtener un conjunto de observaciones T_i^2 homogéneo (ningún punto arriba del UCL).

Aproximación del intervalo de confianza para el percentil $Q_{1-\alpha}$

1. Con el conjunto de datos históricos inicial, calcular los estimadores usuales del vector de medias y de la matriz de varianzas covarianzas. Calcular los valores T_i^2 para las observaciones del conjunto de datos históricos.
2. Mediante la técnica del cuantil, hallar r y s tales que $P\left(T_{(r)}^2 \leq Q_{1-\alpha} \leq T_{(s)}^2\right) \geq \gamma$ sujetos a que $(s - r)$ es mínimo posible, y se obtiene el límite de control como $UCL = \frac{T_{(r)}^2 + T_{(s)}^2}{2}$, de acuerdo con la ecuación (4.75).
3. Comparar los valores T_i^2 con el anterior UCL estimado. Para aquellos valores para los cuales $T_i^2 > UCL$, determinar si existen causas asignables removibles, y eliminar las observaciones correspondientes del conjunto de datos.
4. Con las observaciones que queden, recalcular los estimadores muestrales del vector de medias, la matriz de varianzas covarianzas y calcular de nuevo los valores T_i^2 . Repetir 2 a 4 hasta obtener un conjunto de observaciones T_i^2 homogéneo. Sin embargo, con este método siempre habrá al menos un valor excediendo el UCL, así que se debe parar cuando solo un *outlier* sea encontrado.

Aproximación con la técnica de suavizamiento kernel

1. Estimar el vector de medias y la matriz de varianzas covarianzas del conjunto inicial de observaciones del proceso. Calcular los valores T_i^2 para las observaciones del conjunto de datos históricos.

2. Utilizando el suavizamiento kernel (ecuación (4.76)) sobre la distribución de los valores T_i^2 , estimar el UCL como el percentil $(1 - \alpha)100\%$ de la distribución ajustada, asumiendo que la correlación entre los valores T_i^2 tiene poco efecto en la aplicación de este procedimiento.
3. Comparar los valores T_i^2 *vs.* el UCL hallado en el paso previo. Eliminar del conjunto de datos históricos aquellas observaciones donde $T_i^2 > UCL$ y se halle una causa asignable removible. Tener en cuenta que generalmente la distribución kernel tiende a ajustar bien a los datos para valores moderados de α ; por tanto, aproximadamente $n\alpha$ de los valores T_i^2 aparecerán arriba del UCL, siendo puntos donde ocurre falsa alarma.
4. Con el conjunto de datos que se conservan, recalcular los estimadores muestrales del vector de medias, la matriz de varianzas covarianzas y calcular de nuevo los valores T_i^2 . Repetir 2 a 4 hasta obtener un conjunto de observaciones T_i^2 homogéneo. Sin embargo, en cada paso inevitablemente habrá valores arriba del UCL, a menos que $n\alpha$ sea muy pequeño. Así que para determinar cuándo parar, comparar el UCL actual *vs.* el UCL de la etapa anterior, y si no hubo un cambio significativo, entonces el último UCL hallado es el deseado.

En cualquiera de los casos anteriores, se espera que las estimaciones y el UCL finales determinen el rango operativo de las variables del proceso, el cual no es el mismo que se hallaría desde una perspectiva univariada, a menos que las variables sean independientes. Cuando hay correlación, el rango operativo se aumenta.

4.11 Fase II de control

El proceso de monitoreo en esta fase utiliza funciones de probabilidad diferentes a las de la primera fase para definir las regiones de control. Si hay independencia, se usa la carta de control T^2 de manera análoga a la Fase I, excepto por la función de densidad de probabilidad utilizada:

1. Si μ y Σ son conocidas, entonces $T^2 \sim \chi^2_{(p)}$, y de aquí se encuentra el UCL.
2. Si μ y Σ son desconocidas y deben ser estimados de HDS, entonces

$$T^2 \sim \frac{p(n+1)(n-1)}{n(n-p)} F_{(p,n-p)} \quad (4.77)$$

y

$$UCL = \left(\frac{p(n+1)(n-1)}{n(n-p)} \right) F_{(\alpha;p,n-p)} \quad (4.78)$$

Relativo al número de datos comparados simultáneamente con el UCL, en la Fase II se tiene a la vez solo un valor para la gráfica, en tanto que en la Fase I se tiene más de una observación o valor del estadístico T^2 por comparar simultáneamente con el UCL. Y finalmente, los valores T_i^2 en la Fase II son independientes en tanto que en la Fase I no hay independencia.

Obsérvese que después del inicio de la Fase II, por algún tiempo se utilizan los estimadores muestrales del vector de medias y de la matriz de varianzas covarianzas resultantes de la Fase I; es decir, se trabaja bajo el supuesto de parámetros desconocidos. Luego de operar por algún tiempo con las cantidades estimadas, y haber comprobado la estabilidad del proceso, las estimaciones deben ser tratadas como si fuesen los verdaderos valores de los parámetros.

4.11.1 Escogencia de la tasa de falsa alarma

La tasa de falsa alarma en la Fase II (por ejemplo, error tipo I, α) no tiene por qué coincidir con la de la Fase I. Hay casos en los cuales incurrir en error tipo I durante la Fase II no es tan delicado como en la Fase I. Por ejemplo, si se puede retardar la respuesta a una señal hasta que se observen más de un punto arriba del UCL, entonces en ese caso se requiere un α más grande en la Fase II, con lo cual se reduce el riesgo de sobrecontrolar el proceso. Sin embargo, en la Fase I usar un α grande puede producir estimaciones conservadoras de los parámetros.

En todo caso, también es preciso tener en cuenta la frecuencia de muestreo. Esperar más de una señal antes de llevar a cabo alguna acción puede ser tolerable si no se pone en riesgo la operación del proceso.

4.11.2 Reacción a las señales

Una señal, es decir, una observación arriba del UCL, puede ser debida a una alteración del proceso o bien a una falsa alarma; sin embargo, señales múltiples a menudo implican una desviación del proceso. El problema que se enfrenta es el momento de reacción: cuando ocurra una única señal o cuando más de una sea observada. Esto depende de las características mismas del proceso; es posible que en ciertos procesos los problemas puedan corregirse de inmediato o, por el contrario, que se deba esperar hasta la siguiente actividad de mantenimiento programado. También se puede elegir no reaccionar de inmediato a una señal con el fin de reducir el riesgo de sobrecontrolar el proceso. En este caso, se vigila la ocurrencia de tendencias o patrones en la carta y se declara al proceso fuera de control solo cuando un número dado de observaciones caen arriba del UCL. También se puede consultar a los ingenieros del proceso antes de decidir si se lleva a cabo alguna acción.

En síntesis, para determinar cuándo es apropiado declarar al proceso fuera de control, es necesario conocer bien el proceso, sus factores y componentes críticos, sus riesgos y las consecuencias de acciones o decisiones que se tomen frente a estos. Un análisis de modo y efecto de falla o FMEA podría ser una metodología útil en la realización de esta tarea.

4.11.3 Interpretación de patrones en la carta T^2

En esta carta, como en cualquier carta tipo Shewhart, se espera un comportamiento aleatorio; sin embargo, no se puede esperar que sea como el de un ruido blanco, como en el caso univariado, dado que la distribución del estadístico T^2 no presenta simetría alrededor de la media, es una distribución bastante sesgada a la derecha, así que es muy difícil caracterizar condiciones o patrones que en esta carta permitan predecir la ocurrencia de alteraciones en el proceso. Para algunas industrias, la carta tiene un comportamiento típico y propio del estado estable, así que podría usarse para caracterizar el comportamiento del proceso. Cualquier movimiento consistente lejos del patrón característico para la carta en estado estable puede tomarse como las ocurrencias de condiciones precursoras de un estado alterado del proceso.

Es importante que la influencia de factores, como las condiciones ambientales, sea removida de modo que se pueda producir una imagen correcta del proceso en estado estable y de las relaciones de las variables que lo integran.

Una vez que los patrones en la carta de control T^2 para el proceso en estado estable se han establecido y estudiado, cualquier desviación de dicho patrón es indicador de algún tipo de desviación o cambio, que debe ser investigado para determinar si mejoras o desmejoras están ocurriendo y cuáles son sus causas.

4.12 Control mediante componentes principales

En industrias como la química, petroquímica y farmacéutica no es extraño el monitoreo simultáneo sobre 10, 20 o más variables. En estos casos, puede ser útil emplear una técnica de reducción de dimensionalidad que permita a la vez mantener sensibilidad a la presencia de causas especiales de variación en tales procesos. Una técnica que permite tal aplicación es el análisis de componentes principales.

Considere inicialmente un proceso bivariado de parámetros conocidos:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Sea un vector de observaciones $\mathbf{x} = (x_1, x_2)^t$. El estadístico T^2 definido como $T^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ se puede escribir como $T^2 = \mathbf{z}^t \mathbf{R}^{-1} \mathbf{z}$, donde $\mathbf{R}$ es la matriz de correlación para $\mathbf{x}$ y $\mathbf{z}$ es el vector de componentes estandarizadas $z_i = (x_i - \mu_i)/\sigma_i$. Considere ahora la siguiente transformación:

$$w_1 = \frac{(z_1 + z_2)}{\sqrt{2}} \quad \text{y} \quad w_2 = \frac{(z_1 - z_2)}{\sqrt{2}}$$

entonces, también se cumple que

$$T^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{w}^t \mathbf{L}^{-1} \mathbf{w}$$

donde $\mathbf{w} = (w_1, w_2)^t$ y $\mathbf{L}$ es la matriz diagonal de valores propios de $\mathbf{R}$, con $\lambda_1 \geq \lambda_2$. Esta última transformación define una rotación del espacio de (x_1, x_2) al espacio de (w_1, w_2) , con la cual se ha removido la dependencia, es decir, w_1 y w_2 definen dos ejes ortogonales y constituyen las dos componentes principales de la matriz de correlación de x_1 y x_2 .

La región de control para el estadístico T^2 se puede escribir en términos de w_1 y w_2 de la siguiente manera:

$$T^2 = \frac{w_1^2}{\lambda_1} + \frac{w_2^2}{\lambda_2} \leq UCL$$

El anterior resultado se puede generalizar al caso de $p > 2$:

$$T^2 = \frac{w_1^2}{\lambda_1} + \frac{w_2^2}{\lambda_2} + \cdots + \frac{w_p^2}{\lambda_p} \leq UCL$$

con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Cada componente principal puede ser calculada como $w_i = \mathbf{v}_{(i)}^t \mathbf{z}$, donde $\mathbf{v}_{(i)}$ es el i -ésimo vector propio normalizado de la matriz de correlación $\mathbf{R}$.

En términos de los datos muestrales, se deben estimar las componentes principales y los correspondientes valores propios. Debido a que los coeficientes de las combinaciones lineales en $\mathbf{v}_{(i)}$ son también estimados, las componentes principales no tienen una distribución normal, aun cuando la población sea normal. Por tanto, es usual recurrir a la aproximación de muestras grandes. A continuación, la región de control se define en función de una distribución chi cuadrado:

Sea una muestra aleatoria $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ de una distribución normal multivariada $N_p(\mu, \Sigma)$. Considere las dos primeras componentes principales muestrales de la matriz de correlación $\hat{w}_{j1} = \hat{\mathbf{v}}_{(1)}^t \hat{\mathbf{z}}_j$ y $\hat{w}_{j2} = \hat{\mathbf{v}}_{(2)}^t \hat{\mathbf{z}}_j$, donde $\hat{\mathbf{z}}_j$ es la j -ésima fila de la matriz de diseño con variables estandarizadas y $\hat{\mathbf{v}}_{(1)}, \hat{\mathbf{v}}_{(2)}$, los vectores propios normalizados muestrales de la matriz de correlación muestral. Se consideran solo las dos primeras componentes puesto que explican la proporción acumulada más grande de la varianza muestral total, la cual es igual a $\sum_{i=1}^p \hat{\lambda}_i$, con $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$, los valores propios estimados de la matriz de correlación muestral.

Si el proceso es estable en el tiempo, entonces los valores de las dos primeras componentes principales también deberían serlo, y viceversa. El

proceso de monitoreo de la calidad usando las componentes principales consta de dos partes, a saber:

1. Monitorear las dos primeras componentes principales en cada observación $(\hat{w}_{j1}, \hat{w}_{j2})$ y determinar si existe alguna de ellas por fuera de la región de control, y la variable o variables que aportan en mayor medida a dicha señal. La región de control establecida por las dos primeras componentes principales está definida por la elipse dada por

$$\frac{\hat{w}_1^2}{\hat{\lambda}_1} + \frac{\hat{w}_2^2}{\hat{\lambda}_2} \leq \chi_{2,\alpha}^2$$

2. Monitorear las $p-2$ últimas componentes mediante carta de control T^2 definida por estas para establecer si persisten señales de fuera de control no detectadas por las dos primeras componentes. En este caso el estadístico de monitoreo es

$$T_j^2 = \frac{\hat{w}_3^2}{\hat{\lambda}_3} + \frac{\hat{w}_4^2}{\hat{\lambda}_4} + \cdots + \frac{\hat{w}_p^2}{\hat{\lambda}_p}, \quad j = 1, 2, \dots, n$$

y el UCL de la carta está dado por $\chi_{p-2,\alpha}^2$. En caso de una señal hay que identificar cuál componente, y dentro de esta, cuál o cuáles variables la producen.

Nota. El ejemplo que viene a continuación solo sirve para ilustrar el uso del PCA en el monitoreo de la calidad; sin embargo, para las aproximaciones propuestas se requiere un conjunto de datos “suficientemente grande”).

Considere el conjunto de datos históricos de la tabla 4.5 referentes a datos sobre distintas clases de horas extras en la jornada de trabajo en cierto Departamento de Policía, donde cada observación representa un total de 12 periodos de pago. Las variables se refieren a:

- X1: horas extras por audiencias legales
- X2: horas extras por ocurrencia de eventos extraordinarios
- X3: horas extras por controles
- X4: horas extras por compensatorios autorizados
- X5: horas extras por reuniones

TABLA 4.5 Cinco tipos de horas extras para el Departamento de Policía, Madison, Wisconsin.

obs	X1	X2	X3	X4	X5
1	3387	2200	1181	14861	236
2	3109	875	3532	11367	310
3	2670	957	2502	13329	1182
4	3125	1758	4510	12328	1208
5	3469	868	3032	12847	1385
6	3120	398	2130	13979	1053
7	3671	1603	1982	13528	1046
8	4531	523	4675	12699	1100
9	3678	2034	2354	13534	1349
10	3238	1136	4606	11609	1150
11	3135	5326	3044	14189	1216
12	5271	1658	3340	15052	660
13	3728	1945	2111	12236	299
14	3506	344	1291	15482	206
15	3824	807	1365	14900	239
16	3516	1223	1175	15078	161

Fuente: Johnson & Wichern (1998), p. 258.

En primer lugar, para ilustrar el primer paso del monitoreo de calidad presentado en la página ??, se calcularán las componentes principales sobre la matriz de correlación muestral (es decir, usando los datos estandarizados). Para ello, se puede usar el siguiente programa en R:

Código R 4.27

```
# Creación de la matriz de Diseño X:
X<-matrix(scan(),ncol=5,byrow=T)
3387 2200 1181 14861 236 3109 875 3532 11367 310
2670 957 2502 13329 1182 3125 1758 4510 12328 1208
3469 868 3032 12847 1385 3120 398 2130 13979 1053
3671 1603 1982 13528 1046 4531 523 4675 12699 1100
3678 2034 2354 13534 1349 3238 1136 4606 11609 1150
3135 5326 3044 14189 1216 5271 1658 3340 15052 660
3728 1945 2111 12236 299 3506 344 1291 15482 206
3824 807 1365 14900 239 3516 1223 1175 15078 161

# Cálculo de las componentes principales:
(comp.princ<-prcomp(X, retx=TRUE, center=TRUE, scale.=TRUE,
tol=NULL))
```

Salida R 4.23

```
# Las desviaciones estándar de las componentes principales,
#es decir, raíz cuadrada de los respectivos valores propios:
Standard deviations:
```

```
[1] 1.4707248 1.0696157 1.0135339 0.7227421 0.3785285
```

4.12. CONTROL MEDIANTE COMPONENTES PRINCIPALES

153

#Los vectores propios (los loadings):

Rotation:

	PC1	PC2	PC3	PC4	PC5
[1,]	0.15636908	-0.6510652	0.6569626	0.06472755	0.3404017
[2,]	-0.07330454	0.6740633	0.6103725	0.40381769	0.0681306
[3,]	-0.59922033	-0.2881410	0.1983501	0.18959833	-0.6947082
[4,]	0.57680293	0.1112941	0.3192487	-0.46581653	-0.5796613
[5,]	-0.52764433	0.1622993	0.2336304	-0.76145194	0.2466945

#Obtención de la tabla de varianzas de las componentes:

summary(comp.princ)

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.471	1.070	1.014	0.723	0.3785
Proportion of Variance	0.433	0.229	0.205	0.104	0.0287
Cumulative Proportion	0.433	0.661	0.867	0.971	1.0000

#Los valores de las cinco componentes principales

(compon<-comp.princ\$x)

	PC1	PC2	PC3	PC4	PC5
[1,]	1.85531216	0.86970048	-0.01903857	0.43543350	-0.06469305
[2,]	-0.92840682	-0.42575271	-1.43239519	1.45300361	-0.05425506
[3,]	-0.63599853	0.79647490	-1.11669556	-0.82878030	-0.11836038
[4,]	-2.03851444	0.21673578	-0.12106759	0.12373225	-0.51726476
[5,]	-1.13006584	-0.19621946	-0.24209992	-0.84585484	0.33287902
[6,]	0.12864559	0.10274370	-0.89102727	-1.05590115	-0.04433682
[7,]	0.07534145	0.20191003	0.17976193	-0.43743931	0.61142668
[8,]	-1.40277948	-2.01667018	0.80478340	-0.08349922	-0.12815304
[9,]	-0.46877217	0.45563233	0.62153064	-0.72000872	0.58114519
[10,]	-2.27261018	-0.36123952	-0.51056862	0.28848022	-0.25628925
[11,]	-0.71493912	2.75094432	1.94986261	0.43471350	-0.28654316
[12,]	0.90817111	-1.78258875	2.31942358	0.03781890	-0.16180860
[13,]	0.26421562	-0.06025848	-0.24648300	1.36726915	0.77433504
[14,]	2.25419459	-0.29716068	-0.69456734	-0.34223008	-0.46099606
[15,]	1.97511993	-0.42522170	-0.23090349	0.01587044	-0.02524992
[16,]	2.13108613	0.17096993	-0.37051561	0.15739204	-0.18183583

#Obtención de los valores propios:

(val.eigen<-comp.princ[[1]]^2)

[1] 2.1630313 1.1440778 1.0272509 0.5223561 0.1432838

#Obtención de las dos primeras componentes:

(dos.princomp<- comp.princ\$x[,1:2])

	PC1	PC2
[1,]	1.85531216	0.86970048
[2,]	-0.92840682	-0.42575271
[3,]	-0.63599853	0.79647490
[4,]	-2.03851444	0.21673578
[5,]	-1.13006584	-0.19621946
[6,]	0.12864559	0.10274370
[7,]	0.07534145	0.20191003

```
[8,] -1.40277948 -2.01667018
[9,] -0.46877217  0.45563233
[10,] -2.27261018 -0.36123952
[11,] -0.71493912  2.75094432
[12,]  0.90817111 -1.78258875
[13,]  0.26421562 -0.06025848
[14,]  2.25419459 -0.29716068
[15,]  1.97511993 -0.42522170
[16,]  2.13108613  0.17096993
```

Se puede observar que entre las dos primeras componentes apenas se explica el 66,1 % de la variabilidad total, luego hay todavía información importante sobre la variabilidad del proceso en las últimas tres componentes.

A continuación se construye la carta elipse para las dos primeras componentes, usando un nivel de significancia de 0,05. En este caso la región de control está definida por el cuantil $\chi^2_{0,05,2} = 5,991465$. Para la construcción de la carta elipse, se puede emplear el siguiente programa en R (ver Correa & González, 2002, p. 221):

Código R 4.28

```
carta.elipse<-function(PC1,PC2, val.prop1, val.prop2, k=1000,
alpha=0.05){A<-diag(c(val.prop1, val.prop2), ncol=2)
m<-c(0,0)
const<-sqrt(qchisq(alpha, 2, ncp=0, lower.tail = FALSE))
r<-A[1, 2]/sqrt(A[1, 1] * A[2, 2])

# Construye una matriz nula Q
Q<-matrix(0, 2, 2)

# Transformacion del circulo unitario a una elipse
Q[1, 1]<-sqrt(A[1, 1] %*% (1+r)/2)
Q[1, 2]<- -sqrt(A[1, 1] %*% (1-r)/2)
Q[2, 1]<-sqrt(A[2, 2] %*% (1+r)/2)
Q[2, 2]<-sqrt(A[2, 2] %*% (1-r)/2)

# Define ángulos para graficar
theta<-seq(0, by = (2 * pi)/k, length = k)

# Define coordenadas de puntos sobre círculo unitario
Z<-cbind(cos(theta), sin(theta))

# Define coordenadas de puntos sobre la elipse
Y <- t(m + const * Q %*% t(Z))
plot(Y[,1],Y[,2],type='l', col="orange", lwd=2.0, xlab='comp.1',
ylab='comp.2', xlim=c(min(Y[,1], PC1)-0.5, max(Y[,1], PC1)+0.5),
ylim=c(min(Y[,2], PC2)-0.5, max(Y[,2], PC2)+0.5),
```

```
main=paste("Carta ellipse comp. 1 vs. comp. 2,",sep=" ",
"alpha =",alpha),cex.main=0.9)n<-length(PC1)
for(i in 1:n){
temp<-ifelse(((PC1[i])^2/val.prop1+(PC2[i])^2/val.prop2)>=
const^2,17,19)
temp2<-ifelse(((PC1[i])^2/val.prop1+(PC2[i])^2/val.prop2)>=
const^2,"red","darkblue"); points(PC1[i],PC2[i],pch=temp,
col=temp2,cex=1.2)
if(((PC1[i])^2/val.prop1+(PC2[i])^2/val.prop2)>=const^2)
text(PC1[i],PC2[i],labels=paste('obs =',i),pos=3,font=2,
cex=0.7)}
points(0,0,pch="+",cex=1.5)
abline(h=0,lty=2)
abline(v=0,lty=2) }
```

#Aplicación de la función carta.ellipse a los resultados del PCA: carta.ellipse(dos.princomp[,1],dos.princomp[,2],val.eigen[1],val.eigen[2])

Salida R 4.24

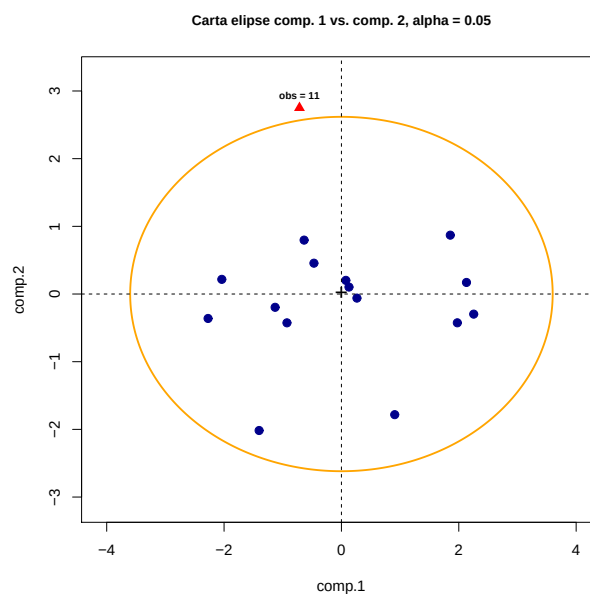


FIGURA 4.21 Carta ellipse al 5 % de significancia, para las dos primeras componentes, datos de la tabla 4.5. La observación 11 aparece fuera de control en la segunda componente.

Con el código R 4.28 resulta el gráfico de la figura 4.21. Como se puede ver, la observación 11 se encuentra fuera de la región de control, siendo responsable de esto la segunda componente. Esto se puede verificar examinando los valores de los *scores* en dicha observación, la cual posee para la segunda componente el valor (en términos absolutos) más grande entre las 16 observaciones. Se debe indagar entonces cuál (o cuáles) de las cinco variables es causante de dicha señal.

Si se inspeccionan los *loadings* de la segunda componente, las variables estandarizadas Z_1 y Z_2 tienen los dos mayores coeficientes, $-0,6510652$ y $0,6740633$, respectivamente. Sin embargo, dado que el *score* en la segunda componente de la observación 11 fue significativamente “grande” (positivo), este valor debe atribuirse o a un valor significativamente negativo en Z_1 o a un valor significativamente positivo en Z_2 , o ambos casos. Para decidir, hay que evaluar la matriz de datos estandarizados. A continuación se presenta un procedimiento en R con el cual se obtienen los datos estandarizados:

Salida R 4.25

Estandarización de los datos:

(Z<-scale(X))

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-0.28246300	0.6100380	-1.2386131	0.99553195	-1.1897622
[2,]	-0.73343049	-0.5101703	0.7080599	-1.68556537	-1.0336587
[3,]	-1.44556980	-0.4408442	-0.1447998	-0.18003791	0.8058319
[4,]	-0.70747553	0.2363534	1.5178626	-0.94814850	0.8606791
[5,]	-0.14944381	-0.5160884	0.2940503	-0.54989736	1.2340619
[6,]	-0.71558645	-0.9134453	-0.4528230	0.31873520	0.5337054
[7,]	0.17823760	0.1053101	-0.5753698	-0.02733660	0.5189388
[8,]	1.57331689	-0.8077653	1.6544857	-0.66346416	0.6328522
[9,]	0.18959290	0.4696949	-0.2673467	-0.02273254	1.1581196
[10,]	-0.52416860	-0.2895104	1.5973524	-1.49986830	0.7383276
[11,]	-0.69125368	3.2528841	0.3039865	0.47987729	0.8775551
[12,]	2.77373395	0.1518094	0.5490802	1.14209451	-0.2953310
[13,]	0.27070216	0.3944507	-0.4685553	-1.01874408	-1.0568632
[14,]	-0.08942296	-0.9590991	-1.1475310	1.47205211	-1.2530474
[15,]	0.42643194	-0.5676603	-1.0862576	1.02545834	-1.1834337
[16,]	-0.07320111	-0.2159571	-1.2435812	1.16204544	-1.3479753

attr(,"scaled:center")

[1] 3561.125 1478.438 2676.875 13563.625 800.000

attr(,"scaled:scale")

[1] 616.4524 1182.8157 1207.7016 1303.1978 474.0443

Se puede ver del anterior resultado que, en la observación 11, la variable Z_2 tiene un valor de 3,2528841, mucho mayor que en el resto de observaciones, en tanto que la variable Z_1 toma un valor de $-0,69125368$ relativamente cercano al del resto de observaciones en dicha variable. Por tanto, la señal se debe a la variable original X_2 , bien sea por una desviación en la media o en su varianza; para verificarlo, se puede realizar una carta de control univariada (carta X en este caso).

Por otra parte, para ilustrar el segundo paso del monitoreo de calidad, se observa que como cerca del 34 % de la variabilidad total del proceso es explicada por el resto de componentes principales, a continuación se verifica si no existen situaciones adicionales de fuera de control que alcancen a ser detectadas mediante estas componentes. Para ello, se construye la carta T^2 para las últimas tres componentes, con los T_j^2 contruidos según numeral 2 del procedimiento descrito en las páginas 150-151. El UCL de la carta se aproxima con el cuantil $\chi_{0,05,3}^2 = 7,814728$.

Código R 4.29

#Carta T2 para la variabilidad no explicada por dos primeras componentes

```
cartaT2.resto.comp<-function(compon,val.eigen,p,alpha=0.05){
  resto.comp<-compon[, -1:-2]
  n<-nrow(resto.comp)
  resto.val<-matrix(rep(val.eigen[-1:-2],n),ncol=3,byrow=T)
  UCL<-qchisq(alpha,df=p-2,lower.tail=FALSE)
  T2<-apply(((resto.comp^2)/resto.val),1,sum)
  Observacion<-1:n
  plot(Observacion,T2,type='l',col="orange",lwd=2,xlim=c(0,n+2),
  ylim=c(0,max(UCL,max(T2))+2),main=expression("Carta"*T^2*~
  "Últimas componentes principales"),ylab=expression(T^2),
  xlab="No. Observación",font=2,font.lab=2)
  abline(h=UCL,lty=3,lwd=2.0,col="darkblue")
  for(i in 1:n){
    temp<-ifelse((T2[i]>UCL),17,19)
    temp2<-ifelse((T2[i]>UCL),"red","darkblue")
    points(Observacion[i],T2[i],pch=temp,col=temp2,cex=1.2)
    if(T2[i]>UCL)text(i,T2[i],labels=paste('obs=',i),pos=3,font=2,
    cex=0.7) }
  text((max(Observacion)+0.5),UCL,paste('UCL=',round(UCL,
  digits=4)),pos=3,font=2,cex=0.7)
  legend(locator(1),c(paste("p=",p),paste("alpha=",alpha),
  paste("n=",n)),ncol=3,cex=0.7,bg='gray95')
}
```

#Aplicación de la función cartaT2.resto.comp a los resultados

del PCA:

#Hacer clic en región del gráfico donde se desea la leyenda

```
cartaT2.resto.comp(compon, val.eigen, 5)
```

Con lo anterior se obtiene la gráfica que aparece en la figura 4.22.

Salida R 4.26

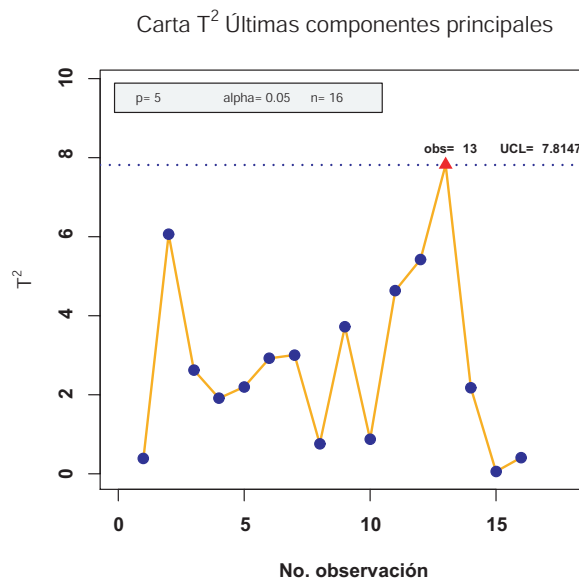


FIGURA 4.22 Carta de control T^2 de las últimas tres componentes, datos de la tabla 4.5. La observación 13 aparece fuera de control.

La observación 13 aparece arriba del UCL; por tanto es necesario evaluar cuál de las tres componentes explica dicha señal y, dentro de esta, cuál variable (o variables) es responsable. Note que, aparentemente, la perturbación del proceso detectado en el periodo 11 continúa por lo menos hasta el periodo 13; observe la tendencia hacia arriba en la gráfica entre las observaciones 11 a 13. Examinando los *scores* de las tres últimas componentes en el periodo 13, se concluye que la componente cuatro puede ser la responsable de la señal pues toma un valor relativamente grande (1,36726915) comparativamente con el resto de periodos. En esta componente, las variables con mayor peso (en términos absolutos) son Z_2 , Z_4 y Z_5 , con *loadings* de 0,40381769, $-0,46581653$ y $-0,76145194$, respectivamente. De estas tres variables parece que Z_4 y Z_5 pueden ser las

causantes de la señal. Sin embargo, si se evalúa la situación en términos de las variables originales, se puede ver que en el periodo 13 la variable X_5 tomó uno de los valores más pequeños (299), aparentemente debido al notable incremento de la variable X_2 en el periodo 11.

En conclusión, en el periodo 11 ocurren eventos extraordinarios que demandan más horas extras, y el efecto de este evento sobre el sistema de causas de horas extras analizado es que las horas dedicadas a reuniones y por compensatorios autorizados deban reducirse en los dos periodos siguientes.

Si se considera que la observación del periodo 11 está fuera de control por causas asignables, se debe eliminar dicha observación del conjunto de datos históricos, para recalculer las componentes principales y la carta elipse para las dos primeras componentes; si nuevos puntos permanecen fuera de control con causas asignables, deberán eliminarse de la base de datos; se continúa hasta obtener un conjunto en control.

Usualmente 50 o más observaciones son necesarias para estimar las componentes principales del proceso en control y calcular los UCL. Una vez se tenga el HDS, se puede usar el esquema de monitoreo con componentes principales, sobre observaciones futuras, donde básicamente el límite de control sigue siendo aproximado con la distribución χ^2 antes vista. El nivel de error tipo I puede ser inferior al manejado previamente en la depuración de los datos históricos.

Finalmente, si el número de variables por controlar es grande, es muy posible que las últimas componentes principales estén asociadas a valores propios muy pequeños, lo cual inflará indebidamente el estadístico T^2 en la carta T^2 de las últimas $p-2$ componentes principales. Una alternativa (Johnson & Wichern, 1998) consiste en:

1. Para cada observación estable, calcular la suma de cuadrados de las últimas $p-2$ componentes

$$d_{Uj}^2 = \sum_{k=3}^p \hat{w}_{jk}^2$$

2. Graficar la carta de control d_{Uj}^2 vs. j . El UCL de la carta es aproximadamente $c\chi_{\alpha,\nu}^2$, donde

$$\begin{aligned}\overline{d_{Uj}^2} &= \frac{1}{n} \sum_{j=1}^n d_{Uj}^2 = c\nu \\ S_{d_{uj}^2}^2 &= \frac{1}{n-1} \sum_{j=1}^n \left(d_{Uj}^2 - \overline{d_{Uj}^2} \right)^2 = 2c^2\nu \\ c &= \frac{S_{d_{uj}^2}^2}{2\overline{d_{Uj}^2}}, \quad \nu = 2 \frac{\left(\overline{d_{Uj}^2} \right)^2}{S_{d_{uj}^2}^2}\end{aligned}$$

4.13 Interpretación de señales

Cuando surge una señal en la carta de control T^2 , es necesario evaluar:

- Cuál o cuáles variables han resultado perturbadas.
- Cuáles parámetros se han desviado: medias, varianzas, covarianzas.

Una señal puede surgir porque:

- alguna(s) variable(s) están por fuera de su rango operativo establecido por el conjunto de datos históricos de base o HDS, o
- El valor de alguna(s) variable(s) no corresponde a la estructura de correlación lineal establecida por el HDS, o
- Por una combinación de los dos casos previos.

A continuación se presenta la técnica de descomposición MYT propuesta por Mason *et al.* (1995), que se aplica en la Fase II para monitorear observaciones individuales.

4.13.1 Descomposición MYT, caso bivariado

La descomposición MYT es una técnica de interpretación de señales, basada en la descomposición del estadístico T^2 en componentes ortogonales que se clasifican en dos tipos:

- Componentes incondicionales: debidas a las variables consideradas individualmente. En caso de aportar a la señal informaría que la variable está por fuera del rango operativo delimitado por el HDS.
- Componentes condicionales: debidas a asociaciones lineales entre subgrupos de variables que, de resultar significativas, estarían indicando que el grupo de variables asociadas a dichas componentes presentan perturbación en su estructura de correlación.

Considere, en primer lugar, el caso de monitoreo de un proceso normal bivariado. Sea $\mathbf{x} = (x_1, x_2)^t$ un vector cuyo estadístico T^2 es:

$$\begin{aligned} T^2 &= (\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{S}_u^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \\ &= \frac{1}{1 - r^2} \left[\frac{(x_1 - \bar{x}_1)^2}{s_1^2} - 2r \frac{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{s_1 s_2} + \frac{(x_2 - \bar{x}_2)^2}{s_2^2} \right] \end{aligned}$$

donde $\bar{\mathbf{x}}$, $\mathbf{S}_u$ son las estimaciones del vector de medias y la matriz de varianzas covarianzas obtenidas a partir del HDS de tamaño n , y r es la correlación entre X_1 y X_2 determinada igualmente por el HDS (en la Fase I). Se puede reescribir la última ecuación de la siguiente manera:

$$T^2 = \frac{1}{1 - r^2} (y_1^2 - 2ry_1y_2 + y_2^2)$$

con $y_i = (x_i - \bar{x}_i)/s_i$. De esta ecuación es evidente que el estadístico T^2 determina una elipse, inclinada. La transformación de componentes principales es una transformación ortogonal que elimina el producto cruzado y_1y_2 , y expresa al estadístico T^2 como la suma de dos componentes aditivas e independientes,

$$T^2 = \frac{\hat{w}_1^2}{\hat{\lambda}_1} + \frac{\hat{w}_2^2}{\hat{\lambda}_2}$$

Sin embargo, w_1 y w_2 son una combinación lineal de ambas variables, lo cual impide una interpretación clara del origen de la señal. Por ello,

se presenta a continuación la descomposición MYT, que es una descomposición ortogonal que permite interpretar las componentes individuales. Dicha descomposición está dada por:

$$T^2 = T_1^2 + T_{2|1}^2 \quad (4.79)$$

o bien

$$T^2 = T_2^2 + T_{1|2}^2 \quad (4.80)$$

donde

$$T_j^2 = \frac{(x_j - \bar{x}_j)^2}{s_j^2}$$

$$T_{i|j}^2 = \frac{(x_i - \bar{x}_{i|j})^2}{s_{i|j}^2}$$

La justificación de esta descomposición se puede ver en forma general en el libro de Johnson & Wichern (1998, p. 218). $i, j = 1, 2$. T_j^2 es la componente incondicional atribuible a la variable X_j , en tanto que $T_{i|j}^2$ es la componente condicional debida a la regresión lineal de la variable X_i en la variable X_j , donde

$$\bar{x}_{i|j} = \hat{E}(X_i | X_j) = \bar{x}_i + \hat{\beta}(x_j - \bar{x}_j)$$

es la media condicional estimada de $X_i | X_j$, es decir la respuesta predicha para X_i al utilizar la ecuación de regresión de X_i sobre X_j , estimada con base en el HDS.

$$s_{i|j}^2 = \hat{V}(X_i | X_j) = \frac{n-2}{n-1} MSE_{i|j}$$

es la varianza condicional estimada (un estimador sesgado) en la regresión de X_i sobre X_j , donde $MSE_{i|j}$ es el error cuadrático medio de dicha regresión.

Recuerde que, bajo normalidad bivariada, la distribución condicional de X_i dado X_j es:

$$f(x_i | x_j) = \frac{1}{\sqrt{2\pi}\sigma_{x_i|x_j}} \exp \left[-\frac{1}{2} \frac{(x_i - \mu_{x_i|x_j})^2}{\sigma_{x_i|x_j}^2} \right]$$

$$\mu_{x_i|x_j} = \mu_{x_i} + \beta(x_j - \mu_{x_j})$$

$$\beta = \rho_{x_i, x_j} \sigma_{x_i} / \sigma_{x_j}$$

$$\sigma_{x_i|x_j} = \sigma_{x_i} \sqrt{(1 - \rho_{x_i, x_j}^2)}$$

Todas las anteriores estimaciones son obtenidas del HDS. Si se denota el residual $x_i - \bar{x}_{i|j}$ por $e_{i|j}$ y el coeficiente de determinación muestral de la regresión por $R_{i|j}^2$, se tiene que la componente condicional se puede expresar como

$$T_{i|j}^2 = \frac{(e_{i|j}/s_i)^2}{1 - R_{i|j}^2}$$

entonces la componente condicional $T_{i|j}^2$ explica qué tan bien una observación futura sobre la variable X_i concuerda con el valor predicho por la nueva observación de la variable X_j con base en el HDS. Esta componente puede ser grande por una correlación cercana a 1 o porque el residual estandarizado $e_{i|j}/s_i$ es muy grande; en dicho caso, no hay concordancia entre el valor observado y el valor predicho de X_i dado el valor observado de X_j .

Distribución de probabilidad de las componentes de la descomposición MYT

Para determinar cuándo alguna de las componentes de la descomposición MYT es estadísticamente grande, se debe fijar un límite de control para cada una, con base en la distribución de probabilidad de estas:

$$T_j^2 \sim \left(\frac{n+1}{n} \right) F_{1, n-1}, \quad j = 1, 2 \quad (4.81)$$

$$T_{i|j}^2 \sim \frac{(n+1)(n-1)}{n(n-2)} F_{1,n-2}, \quad i \neq j, \quad i, j = 1, 2 \quad (4.82)$$

Considerando un nivel de significancia α , una muestra histórica de base o HDS de tamaño n , un UCL para cada término de la descomposición está dado por:

- Términos incondicionales:

$$UCL_{(X_j)} = \left(\frac{n+1}{n} \right) F_{(\alpha,1,n-1)} \quad (4.83)$$

- Términos condicionales:

$$UCL_{(X_i|X_j)} = \frac{(n+1)(n-1)}{n(n-2)} F_{(\alpha,1,n-2)} \quad (4.84)$$

Un $T_j^2 > UCL_{(X_j)}$ indica que la correspondiente variable X_j contribuye a la señal por estar fuera de su rango operativo, y un $T_{i|j}^2 > UCL_{(X_i|X_j)}$ muestra que tanto X_1 como X_2 contribuyen a la señal debido a una perturbación de su asociación.

Al fijar los anteriores límites, la región de aceptación definida por T_j^2 y $T_{i|j}^2$ aproxima la región de control elíptica definida para X_1 y X_2 , por un paralelogramo tal como el que se presenta en la figura 4.23 (la región elíptica queda inscrita en este paralelogramo), cuyo tamaño es controlado por la elección de la significancia empleada en ambos UCL. Podría manejarse una significancia distinta para evaluar el T^2 total y los términos de la descomposición MYT; sin embargo, para evitar ambigüedades, se recomienda usar la misma significancia.

La figura 4.24 presenta cuatro puntos, todos con un valor T^2 grande (produjeron señal). Una descomposición MYT arroja los valores de la tabla 4.6, donde (*) denota significancia al 5%.

Se puede observar que el punto A produce señal porque el término incondicional T_2^2 es grande, lo cual indica que la nueva observación sobre X_2 está fuera del rango operativo definido por el HDS. También se observa que A está lejos de la recta de regresión $\bar{X}_{2|1}$; de ahí que el término

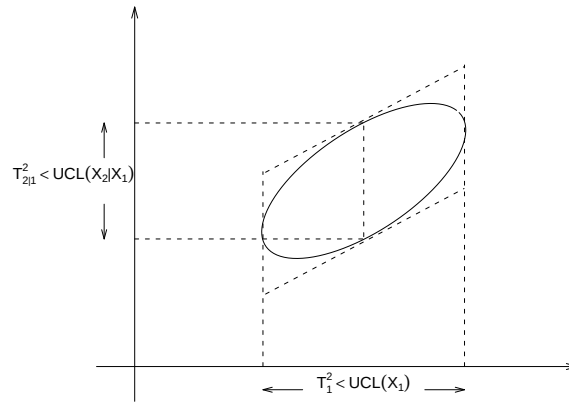


FIGURA 4.23 Región de control elíptica y región de control definida por T_1^2 y $T_{2|1}^2$.

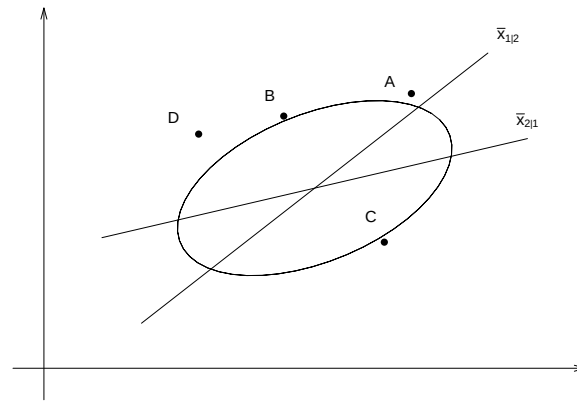


FIGURA 4.24 Región de control elíptica y rectas de regresión relativas a las componentes condicionales $T_{1|2}^2$ y $T_{2|1}^2$.

condicional $T_{2|1}^2$ sea grande, es decir, que la nueva observación sobre X_2 no está en el rango condicional definido por la respectiva observación sobre X_1 , con base en la ecuación de regresión obtenida en el HDS. El punto B también produce señal en el término condicional $T_{2|1}^2$ puesto que también está lejos de la recta de regresión $\bar{X}_{2|1}$, aunque no produce señales en los términos incondicionales dado que este punto se encuentra dentro del rango operativo de X_1 y de X_2 , respectivamente. El punto C está más lejos de la recta de regresión $\bar{X}_{1|2}$, por lo que produce señal en el término condicional, $T_{1|2}^2$, es decir está por fuera del rango condicional

TABLA 4.6 Resumen descomposición MYT para ejemplo bivariado.

punto	T^2	T_1^2	T_2^2	$T_{1 2}^2$	$T_{2 1}^2$
A	10,05*	2,78	10,03*	0,02	7,27*
B	6,33*	0,11	3,49	2,83	6,22*
C	6,63*	3,01	0,34	6,29*	3,62
D	9,76*	2,54	1,73	8,03*	7,22*

* Denota significancia al 5 %.

Fuente: Mason & Young (2002), p. 134.

definido por el valor de X_2 , aunque permanece dentro de los rangos operativos de cada una de las dos variables. En el punto D ambos términos condicionales son significativos, lo cual indica que las observaciones sobre cualquiera de las dos variables en este punto no están donde deberían estar en relación con la posición de la otra variable.

4.13.2 Descomposición MYT, caso general

En el caso general con p variables, considere la partición del vector $(\mathbf{x} - \bar{\mathbf{x}})^t$ en la siguiente forma:

$$(\mathbf{x} - \bar{\mathbf{x}})^t = \left(\mathbf{x}^{(p-1)} - \bar{\mathbf{x}}^{(p-1)}, x_p - \bar{x}_p \right)^t$$

donde

$$\mathbf{x}^{(p-1)} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{p-1} \end{pmatrix}, \quad \bar{\mathbf{x}}^{(p-1)} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_{p-1} \end{pmatrix}$$

por tanto, la matriz de varianzas covarianzas también queda particionada, y es de la siguiente forma:

$$\mathbf{S}_u = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xX} \\ \mathbf{S}_{xX}^t & s_p^2 \end{pmatrix}$$

donde $\mathbf{S}_{xx}$ es la matriz de varianzas covarianzas para las primeras $p - 1$, $\mathbf{S}_{xX}$ es el vector de dimensión $p - 1$ que contiene las covarianzas entre X_p y el resto de $p - 1$ de variables y s_p^2 , la varianza muestral de X_p . Recuerde que los anteriores estimadores se obtienen del conjunto de datos históricos de base o HDS obtenido de la Fase I.

Con base en la anterior partición, se puede demostrar el estadístico T^2 , que puede particionarse de la siguiente manera:

$$T^2 = T_{(x_1, x_2, \dots, x_{p-1})}^2 + T_{p|1,2,\dots,p-1}^2 \quad (4.85)$$

donde

$$T_{(x_1, x_2, \dots, x_{p-1})}^2 = \left(\mathbf{x}^{(p-1)} - \bar{\mathbf{x}}^{(p-1)} \right)^t \mathbf{S}_{xx}^{-1} \left(\mathbf{x}^{(p-1)} - \bar{\mathbf{x}}^{(p-1)} \right)$$

$$T_{p|1,2,\dots,p-1}^2 = \frac{(x_p - \bar{x}_{p|1,2,\dots,p-1})^2}{s_{p|1,2,\dots,p-1}^2}$$

$\bar{x}_{p|1,2,\dots,p-1}$ y $s_{p|1,2,\dots,p-1}^2$ son respectivamente la media (la respuesta predicha) y la varianza condicional estimada de X_p sobre el resto de $p-1$ variables, y obtenidos con base en la regresión de X_p sobre dichas variables en el HDS. Estos dos estimadores se pueden expresar de la siguiente manera:

$$\bar{x}_{p|1,2,\dots,p-1} = \bar{x}_p + \boldsymbol{\beta}_p^t \left(\mathbf{x}^{(p-1)} - \bar{\mathbf{x}}^{(p-1)} \right)$$

donde $\boldsymbol{\beta}_p$ es el vector de coeficientes estimados de la regresión de X_p en el resto de $p-1$ variables (con los datos del HDS), y

$$s_{p|1,2,\dots,p-1}^2 = s_p^2 - \mathbf{S}_{xX}^t \mathbf{S}_{xx}^{-1} \mathbf{S}_{xX}$$

El término incondicional $T_{(x_1, x_2, \dots, x_{p-1})}^2$ es en sí un estadístico T^2 , luego también podría descomponerse de forma similar a (4.85):

$$T_{(x_1, x_2, \dots, x_{p-1})}^2 = T_{(x_1, x_2, \dots, x_{p-2})}^2 + T_{p-1|1,2,\dots,p-2}^2 \quad (4.86)$$

Particionando iterativamente de la forma anterior cada término incondicional que resulta en cada paso, se puede llegar a una de las $p!$ formas posibles de la descomposición MYT del estadístico T^2 , como se muestra a continuación:

$$T^2 = T_1^2 + T_{2|1}^2 + T_{3|1,2}^2 + \dots + T_{p|1,2,\dots,p-1}^2 \quad (4.87)$$

donde los términos $T_j^2 = T_{(x_j)}^2$ corresponden a los términos incondicionales dados por

$$T_j^2 = \frac{(x_j - \bar{x}_j)^2}{s_j^2} \quad (4.88)$$

y los términos condicionales $T_{j|1,2,\dots,j-1}^2$ están dados por

$$T_{j|1,2,\dots,j-1}^2 = \frac{(x_j - \bar{x}_{j|1,2,\dots,j-1})^2}{s_{j|1,2,\dots,j-1}^2} \quad (4.89)$$

Los términos condicionales pueden ser calculados también de la siguiente manera:

$$T_{j|1,2,\dots,j-1}^2 = T_{(x_1,x_2,\dots,x_j)}^2 - T_{(x_1,x_2,\dots,x_{j-1})}^2 \quad (4.90)$$

por ejemplo

$$\begin{aligned} T_{2|1}^2 &= T_{(x_1,x_2)}^2 - T_{(x_1)}^2 \\ T_{3|1,2}^2 &= T_{(x_1,x_2,x_3)}^2 - T_{(x_1,x_2)}^2 \end{aligned}$$

además

$$T_{(x_1,x_2,\dots,x_j)}^2 = (\mathbf{x}^{(j)} - \bar{\mathbf{x}}^{(j)})^t \mathbf{S}_{jj}^{-1} (\mathbf{x}^{(j)} - \bar{\mathbf{x}}^{(j)}) \quad (4.91)$$

donde $\mathbf{x}^{(j)}$ es el vector observación (obtenido en la Fase II) de las primeras j variables, $\bar{\mathbf{x}}^{(j)}$ es el vector de medias muestrales de las primeras j variables en el HDS, y $\mathbf{S}_{jj}$ es la matriz de varianzas covarianzas muestral de las primeras j variables en el HDS. Observe que cuando $j = p$ se tiene el estadístico T^2 total, es decir, $T^2 = T_{(x_1,x_2,\dots,x_p)}^2$.

4.13.3 Propiedades de la descomposición MYT

1. El estadístico $T_{(x_1,x_2,\dots,x_j)}^2$ es invariante bajo permutaciones; por ejemplo,

$$T_{(x_1,x_2,\dots,x_j)}^2 = T_{(x_2,x_1,\dots,x_j)}^2$$

2. Como hay $p!$ posibles permutaciones entre las p variables, se puede particionar el estadístico T^2 en $p!$ formas diferentes, donde cada descomposición corresponde a una permutación. Por ejemplo, si $p = 3$, las $3! = 6$ posibles descomposiciones son:

$$\begin{aligned} T^2 &= T_1^2 + T_{2|1}^2 + T_{3|1,2}^2 \\ T^2 &= T_1^2 + T_{3|1}^2 + T_{2|1,3}^2 \\ T^2 &= T_2^2 + T_{1|2}^2 + T_{3|1,2}^2 \\ T^2 &= T_2^2 + T_{3|2}^2 + T_{1|2,3}^2 \\ T^2 &= T_3^2 + T_{1|3}^2 + T_{2|1,3}^2 \\ T^2 &= T_3^2 + T_{2|3}^2 + T_{1|2,3}^2 \end{aligned}$$

3. Los p términos en cualquiera de las descomposiciones son independientes entre sí, aunque los términos de descomposición a descomposición no lo son necesariamente.
4. Hay $p \times 2^{p-1}$ términos distintos entre las posibles descomposiciones.
5. Si uno o más términos incondicionales aportan a la señal, se debe examinar el valor T^2 del subvector del resto de variables; por ejemplo, si T_1^2 es significativamente grande, entonces se analiza si $T_{(x_2, \dots, x_p)}^2$ es o no grande (significativo). Si este último no es significativo, entonces no es necesario examinar ningún término adicional de la descomposición MYT que involucre a cualquiera de las variables en dicho subvector.
6. Los $p(2^{p-1} - 1)$ términos condicionales únicos contienen los residuales de todas las posibles regresiones lineales de cada variable con respecto al subconjunto de las otras variables.

Identificación de las variables involucradas en una señal

Considere un vector de observación $\mathbf{x}$ para el cual

$$T^2 = T_{(x_1, x_2, \dots, x_p)}^2 > UCL = \frac{p(n+1)(n-1)}{n(n-p)} F_{\alpha, p, n-p}$$

- **Primer método:** Consiste en la aplicación de un esquema iterativo tipo “forward” (hacia adelante), donde la búsqueda termina cuando se alcance el subconjunto de variables que no contribuyen a la señal, así:

1. Construir los estadísticos T^2 para cada variable individual

$$T_j^2 = T_{(x_j)}^2 = \frac{(x_j - \bar{x}_j)^2}{s_j^2}$$

los cuales son comparados con el valor crítico dado por:

$$UCL_{(x_j)} = \left(\frac{n+1}{n} \right) F_{(\alpha, 1, n-1)}$$

Nota: Recuerde que $\bar{x}_j$ y s_j^2 son estimados desde el HDS.

2. Excluir del vector observación las variables X_j cuando $T_j^2 \geq UCL_{(x_j)}$, puesto que dichas observaciones contribuyen a la señal.
3. Para el subvector de observaciones en las variables que no contribuyeron individualmente a la señal, calcular el estadístico T^2 de todos los pares posibles de las variables $T_{(x_i, x_j)}^2$ $i \neq j$, y el valor crítico dado por

$$UCL_{(x_i, x_j)} = \frac{2(n+1)(n-1)}{n(n-2)} F_{(\alpha, 2, n-2)}$$

4. Excluir del subvector inicial de las observaciones de variables, los pares x_i, x_j para los cuales $T_{(x_i, x_j)}^2 \geq UCL_{(x_i, x_j)}$, ya que las variables X_i y X_j contribuyen a la señal.
5. En este paso, con el nuevo subvector de observaciones sobre las variables que no fueron eliminadas en el paso anterior, formar todos los subgrupos de tamaño 3, y calcular $T_{(x_i, x_j, x_k)}^2$; comparar con el valor crítico dado por:

$$UCL_{(x_i, x_j, x_k)} = \frac{3(n+1)(n-1)}{n(n-3)} F_{(\alpha, 3, n-3)}$$

6. Excluir todas las triplas de variables del subvector inicial en el paso anterior para las que $T_{(x_i, x_j, x_k)}^2 \geq UCL_{(x_i, x_j, x_k)}$.
7. Este procedimiento continúa de manera iterativa hasta obtener un grupo de variables que no contribuyan a la señal.

Ejemplo. Para un proceso con $p = 3$, considere el vector observación $\mathbf{x} = (533, 514, 528)^t$. El HDS fue construido con base en $n = 23$ observaciones. Para el vector observación dado se obtuvo $T^2 = 79,9441$. Se usa un $\alpha = 0,01$; por tanto el UCL es de 17,00456 y se genera una señal. Aplicando el procedimiento antes descrito se obtiene:

1. Cálculo de términos incondicionales para cada variable:

$$\begin{aligned} T_1^2 &= 1,3934 \\ T_2^2 &= 0,0641 \\ T_3^2 &= 11,6578 \end{aligned}$$

Como el $UCL_{(x_j)} = 8,290837$, entonces X_3 contribuye a la señal.

2. Cálculo de $T^2_{(x_i, x_j)}$ con el subvector determinado por las variables X_1 y X_2 , es decir para el par (533, 514):

$$T^2_{(x_1, x_2)} = 1,3935$$

Como $UCL_{(x_i, x_j)} = 12,63793$, entonces ni X_1 ni X_2 contribuyen a la señal. Por tanto solo X_3 es responsable de la señal.

El anterior método tiene como desventaja el hecho de que proporciona poca información sobre cómo las componentes del vector observación “señalador” contribuyen a la señal, dado que no examina los términos condicionales.

Nota. Un esquema “backward” de eliminación, similar al anterior esquema, puede también aplicarse para ubicar las variables que contribuyen a la señal.

- **Segundo método:** Se examinan todos los términos posibles de las descomposición MYT de un vector observación señalador, y se determina cuáles de estos términos toman valores “grandes” (mayores que sus respectivos UCL). Cabe recordar que

$$T_j^2 \sim \left(\frac{n+1}{n} \right) F_{1, n-1}$$

$$T_{j|1,2,\dots,j-1}^2 \sim \frac{(n+1)(n-1)}{n(n-k-1)} F_{1, n-k-1}$$

donde k es el número de variables condicionantes.

Para el ejemplo dado en el primer método, se obtuvo que entre los términos incondicionales solo T_3^2 fue significativo al 0,01. Con los términos condicionales: $T_{1|2}^2 = 1,3294$, $T_{1|3}^2 = 28,23$, $T_{2|1}^2 = 0,0001$, $T_{2|3}^2 = 9,5584$, $T_{3|1}^2 = 38,4949$ y $T_{3|2}^2 = 21,1522$, el valor crítico es

$$\frac{(n+1)(n-1)}{n(n-2)} F_{(0,01,1,n-2)} = 8,763485$$

Observe que todos los términos condicionales que contienen a X_3 son significativamente “grandes”. Se podría concluir que algo pasa con X_3 , como en el primer método, pero además con X_1 y X_2 ,

porque estas variables aparecen en términos de la descomposición con valores grandes.

El número de términos distintos de la descomposición que se deben calcular pueden ser inmanejables.

- **Tercer método: esquema computacional:** Este método busca reducir los cálculos a un número razonable cuando la T^2 produce una señal.

1. Calcular los términos incondicionales T_j^2 ; remover del vector observación las variables X_j con $T_j^2 \geq UCL_{(x_j)}$. Estas variables están fuera de control y, por consiguiente, no se requiere evaluar cómo se asocian a las otras variables.
2. Chequear el subvector de las observaciones sobre las variables restantes para determinar si hay señal. Por ejemplo, si T_1^2 y T_3^2 son significativos en el paso anterior, se eliminan del vector observación señalador y se calcula $T_{(x_2, x_4, \dots, x_p)}^2$, el cual se debe comparar con su respectivo UCL dado por

$$UCL_{(x_2, x_4, \dots, x_p)} = \frac{(p-2)(n+1)(n-1)}{n(n-p+2)} F_{(\alpha, p-2, n-p+2)}$$

Si no hay señal en este subvector, se ha ubicado la fuente del problema.

Nota. En general, *no es cierto* que $T_{(x_2, x_4, \dots, x_p)}^2 = T^2 - T_1^2 - T_3^2$. Para hallar $T_{(x_2, x_4, \dots, x_p)}^2$ es necesario calcularlo utilizando una ecuación similar a la dada en (4.91) involucrando solo a las variables $(x_2, x_4, \dots, x_p)$.

3. Si en el subvector del paso anterior se detecta señal, calcular todos los términos condicionales de la forma $T_{i|j}^2$ posibles entre los pares de variables en dicho subvector. Remover de este último los pares de observaciones asociados a las variables X_i, X_j con $T_{i|j}^2$ significativos, porque en tales variables existe algo anormal con las respectivas asociaciones bivariadas. Con el nuevo subvector de observaciones sobre las variables que no son eliminadas, calcular el respectivo T^2 . Por ejemplo, suponga que se detecta que $T_{2|5}^2$ es significativo; se elimina a X_2 y a X_5 , y queda el subvector de observación sobre las variables $(x_4, x_6, x_7, \dots, x_p)$; se calcula $T_{(x_4, x_6, x_7, \dots, x_p)}^2$, el cual se compara con

$$UCL_{(x_4, x_6, x_7, \dots, x_p)} = \frac{(p-4)(n+1)(n-1)}{n(n-p+4)} f_{\alpha, p-4, n-p+4}$$

si no se detecta señal ($T^2_{(x_4, x_6, x_7, \dots, x_p)} < UCL_{(x_4, x_6, x_7, \dots, x_p)}$), entonces la fuente de los problemas yace en las variables individuales detectadas en el paso 1 y en las relaciones bivariadas perturbadas identificadas en el paso presente.

4. Si en el paso previo se detecta señal en el subvector de variables que quedan (en el ejemplo, $T^2_{(x_4, x_6, x_7, \dots, x_p)}$), se calculan a continuación todos los términos condicionales del tipo $T^2_{i|j,k}$ posibles entre las variables en dicho subvector. Remover todas las triplas de variables (x_i, x_j, x_k) para las cuales dicho estadístico produce señal. Chequear el T^2 del subvector de variables que queden. Por ejemplo, suponga que en este paso el término $T^2_{4|6,8}$ resulta significativo; entonces se elimina del subvector de observaciones a las variables X_4 , X_6 y a X_8 , y queda el subvector formado por las observaciones $(x_7, x_9, x_{10}, \dots, x_p)$ para el cual se calcula el estadístico $T^2_{(x_7, x_9, x_{10}, \dots, x_p)}$ y el valor crítico dado por

$$UCL_{(x_7, x_9, x_{10}, \dots, x_p)} = \frac{(p-7)(n+1)(n-1)}{n(n-p+7)} F_{0,01, p-7, n-p+7}$$

5. Continuar con el esquema anterior hasta que no se detecten más señales.

Interpretación de señales en una componente de la T^2

Términos incondicionales

Cuando están en control, se tiene que

$$T_j^2 < \left(\frac{n+1}{n} \right) F_{\alpha, 1, n-1}$$

y como $F_{\alpha, 1, n-1} = \{t_{\alpha/2, n-1}\}^2$, entonces T_j se encuentra en el intervalo

$$-\sqrt{\left(\frac{n+1}{n} \right) t_{\alpha/2, n-1}^2} < T_j < \sqrt{\left(\frac{n+1}{n} \right) t_{\alpha/2, n-1}^2} \quad (4.92)$$

pero

$$T_j = \frac{x_j - \bar{x}_j}{s_j}$$

entonces un intervalo de control para X_j está dado por

$$\bar{x}_j - t_{\alpha/2, n-1} \sqrt{\left(\frac{n+1}{n}\right)} s_j < X_j < \bar{x}_j + t_{\alpha/2, n-1} \sqrt{\left(\frac{n+1}{n}\right)} s_j \quad (4.93)$$

lo cual es equivalente a usar una carta de control univariada tipo Shewhart sobre la variable X_j . Señales en esta carta indican que la variable X_j opera fuera de rango.

Términos condicionales

Bajo control se tiene que

$$T_{j|1,2,\dots,j-1}^2 < \frac{(n+1)(n-1)}{n(n-k-1)} F_{\alpha,1,n-k-1}$$

donde $k = j - 1$. La componente X_j del vector observación $\mathbf{x} = (x_1, x_2, \dots, x_p)^t$ está contenida en la distribución condicional de X_j dados $X_1, X_2, \dots, X_{j-1}$ y cae en la región de control multivariada definida por el hiperelipsoide de confianza $1 - \alpha$. Si el término condicional está fuera de control, la relación de X_j vs. $X_1, X_2, \dots, X_{j-1}$ está en contra de la relación observada en los datos históricos del HDS, o equivalentemente, X_j no está donde debería estar en relación con los correspondientes valores observados de $X_1, X_2, \dots, X_{j-1}$.

En resumen,

1. Para interpretar las componentes de una descomposición MYT, las señales sobre los términos incondicionales T_j^2 implican que la variable asociada está fuera del rango operacional especificado por el HDS. Un vector de observación con este tipo de señal graficará por fuera del hiperparalelepípedo definido por los límites de control dados en (4.93).
2. Las señales sobre observaciones que aparecen dentro del hiperparalelepípedo, pero que tienen términos condicionales $T_{j|1,2,\dots,j-1}^2$ grandes, implican que algo está errado en la relación entre las variables

involucradas en dichos términos. Todas estas variables necesitan ser examinadas para identificar causas. Para intervalos de control simultáneos sobre m variables individuales, usar intervalos de Bonferroni en (4.93), es decir, con el percentil $t_{\alpha/(2m)}$.

3. Para restablecer el proceso a un estado de control cuando ocurre señal en un término condicional, es necesario ajustar la variable que sea más controlable; pero si todas las variables involucradas son controlables, entonces ajustar aquella con término incondicional más grande entre dichas variables.

4.13.4 La regresión lineal como medio para mejorar la interpretación de señales atribuibles a términos condicionales de la descomposición MYT

La expresión del término condicional dada en la ecuación (4.89) se puede reescribir de la siguiente manera:

$$T_{j|1,2,\dots,j-1}^2 = \left[\frac{e_{j|1,2,\dots,j-1}}{s_j \left(1 - R_{j|1,2,\dots,j-1}^2\right)^{1/2}} \right]^2 \quad (4.94)$$

donde $e_{j|1,2,\dots,j-1}$ es el residual o diferencia entre el valor observado sobre X_j y el valor predicho para esta variable por los valores observados sobre las variables $X_1, \dots, X_{j-1}$, con base en la ecuación de regresión ajustada con el conjunto de datos del HDS. Por tanto, los términos condicionales explican qué tan bien una observación futura sobre la variable X_j concuerda con el valor predicho por el conjunto de las variables condicionantes.

Exceptuando los casos donde $R_{j|1,2,\dots,j-1}^2 \approx 1$, un término condicional grande se debe al desacuerdo entre el valor observado de X_j y el predicho por las variables condicionantes en los términos condicionales. Si $T_{j|1,2,\dots,j-1}^2$ involucra a muchas variables, su tamaño está directamente relacionado a la magnitud de los residuales estandarizados resultantes de la predicción de X_j usando a $X_1, \dots, X_{j-1}$.

En conclusión, una señal ocurre cuando una observación sobre una variable o grupo de variables está fuera de control o cuando las obser-

vacaciones sobre un conjunto de variables están en contra de la relación definida por el conjunto de datos históricos de base o HDS.

CAPÍTULO 5

Métodos robustos para el vector de medias

5.1 Introducción

En la Fase I del monitoreo de un conjunto histórico de datos multivariados, la estimación de los parámetros del modelo, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, es de suma importancia. Una de las principales acciones por realizar en esta fase es la identificación de *outliers* multivariados y cambios en los parámetros del modelo, de tal forma que los límites de control que se usen posteriormente en la Fase II sean lo más precisos posible. Cuando el conjunto histórico de datos presenta grados relativamente altos de contaminación, una escogencia inadecuada de los estimadores –y por consiguiente, una construcción inapropiada del límite de control superior de la carta de control T^2 de Hotelling– produce una carta ineficiente en la detección de patrones no aleatorios. Este problema conduce a que se pase a la Fase II con límites poco precisos, construyéndose una carta de control T^2 relativamente ineficiente para monitorear un proceso. En este capítulo se presentan métodos de análisis en la Fase I, asumiendo que los datos que se toman son observaciones individuales multivariadas.

5.2 Estimación de los parámetros y algunos métodos

Usualmente $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ se estiman mediante el vector de medias muestrales y la matriz de varianzas covarianzas muestral. Sin embargo, este método clásico de estimación no produce límites de control apropiados, pues es bien conocido que estos estimadores son altamente sensitivos a la presencia de *outliers* en los datos. Sullivan & Woodall (1996) y Vargas (2003) demostraron que la carta de control T^2 basada en los estimadores usuales $\bar{\mathbf{x}}$ y $\mathbf{S}$ es ineficiente en términos de la probabilidad de señales fuera de control. La probabilidad de señal es el criterio preferido cuando se desea comparar diversos métodos en el análisis de la Fase I. Cuando los datos provienen de un proceso fuera de control, se espera que la probabilidad de señal sea suficientemente grande para que los puntos fuera de control puedan ser identificados y eliminados.

Métodos alternativos de estimación de parámetros han sido propuestos en la literatura a fin de mejorar la eficiencia de carta de control T^2 . Sullivan & Woodal (1996) sugieren estimar $\boldsymbol{\Sigma}$ mediante el vector de diferencias entre observaciones sucesivas. El estimador $\boldsymbol{\Sigma}$ propuesto es:

$$\mathbf{S}_d = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \mathbf{v}_i \mathbf{v}_i^t \quad (5.1)$$

donde $\mathbf{v}_i = x_{i+1} - x_i$, $i = 1, 2, \dots, m-1$.

Este estimador ya había sido sugerido por Holmes & Mergen (1993). Las estadísticas T^2 basadas en $\mathbf{S}_d$ están dadas por:

$$T_{d,i}^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^t \mathbf{S}_d^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \quad i = 1, \dots, m \quad (5.2)$$

donde $\bar{\mathbf{x}}$ es el vector de medias muestral.

Sullivan & Woodall (1996) demostraron que la carta de control T^2 basada en las estadísticas (5.2) detecta cambios en la media del proceso más rápidamente que la carta de control T^2 usual. Sin embargo, no es eficiente cuando se presentan varios *outliers*.

Una buena alternativa consiste en usar estimadores robustos de localización y dispersión en el cálculo de las estadísticas T^2 . El uso de

estimadores robustos en cartas de control univariadas ha probado ser eficiente. Ver, por ejemplo, Langerberg e Iglewicz (1986), Rocke (1989, 1992), Khoo (2004), de Mast & Roes (2004) y Davis & Adams (2005). Aunque el trabajo de estimadores robustos con datos multivariados lleva varios años, ver, por ejemplo, Campbell (1980), Huber (1981), Maronna (1976), Rousseeuw (1984), Rousseeuw & Leroy (1987) y Rosseeuw *et al.* (2004), en control de procesos multivariados su aplicación es reciente. Vargas (2003) propuso construir una carta de control T^2 insertando en las estadísticas T^2 estimadores robustos de alto punto de rompimiento. Se entiende por punto de rompimiento la máxima fracción de observaciones que puede arbitrariamente ser contaminada antes que el estimador colapse. Específicamente, Vargas (2003) propuso la construcción de cartas T^2 usando estimadores de localización y dispersión MVE (elipsoide de volumen mínimo) y MCD (determinante de covarianza mínimo).

El estimador MVE, propuesto inicialmente por Rousseeuw (1984), busca la elipsoide más pequeña que cubra un subconjunto de por lo menos h puntos, donde $h \leq m$. El estimador de dispersión es la matriz de varianzas covarianzas de los puntos de la elipsoide, multiplicada por una constante con el propósito de garantizar consistencia (ver Rousseeuw y Van Zomeren, 1990). El valor usual de h es $(m + p + 1)/2$. De esta forma se alcanza un punto de rompimiento de $(m + p + 1)/2m\%$, el cual converge a 50% cuando $m \rightarrow \infty$. El estimador MCD de Rousseeuw & Van Driessen (1999), propuesto inicialmente por Rousseeuw (1984), es la media y la matriz de varianzas basada en un subconjunto de h puntos, ($h \leq m$), que minimiza el determinante de la matriz de covarianzas. De nuevo $h = (m + p + 1)/2$, de tal forma que el punto de ruptura es $(m + p + 1)/2m$. Las estadísticas T^2 basadas en estos estimadores se denotan de la siguiente manera:

$$T_{MVE,i}^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{MVE})^t \mathbf{S}_{MVE}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{MVE})$$

y

$$T_{MCD,i}^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})^t \mathbf{S}_{MCD}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})$$

para $i = 1 \dots, m$, donde $\bar{\mathbf{x}}_{MVE}$ y $\bar{\mathbf{x}}_{MCD}$ son los estimadores de localización y $\mathbf{S}_{MVE}$ y $\mathbf{S}_{MCD}$ los estimadores de dispersión MVE y MCD, respectivamente. Las estadísticas T^2 usuales se denotan por $T_{u,i}^2$.

El límite de control superior (UCL) de la carta T^2 basada en los

estimadores usuales en la etapa retrospectiva de la Fase I está dado por:

$$UCL = \frac{(m-1)^2}{m} B_{p/2, (m-p-1)/2, \alpha}$$

donde $B_{\delta_1, \delta_2, \alpha}$ es el percentil $(1 - \alpha)$ de la distribución beta con parámetros δ_1 y δ_2 . Sin embargo, cuando se usan otros estimadores, la distribución exacta de la estadística T^2 no es conocida. Se debe entonces utilizar alguna aproximación asintótica, o calcularlo mediante simulación de Monte Carlo.

Sullivan & Woodall (1996) propusieron la siguiente aproximación para la estadística T^2 basada en $\mathbf{S}_d$:

$$\frac{m}{(m-1)^2} T_{d,i}^2 \sim B_{p/2, (f-p-1)/2}$$

donde $f = 2(m-1)^2/(3m-4)$. Williams *et al.* (2006) muestran que esta aproximación no es apropiada para valores pequeños de m . Como la distribución asintótica de $T_{d,i}^2$ es χ_p^2 para valores grandes de m , Williams *et al.* (2006), a través de resultados de simulaciones, sugieren construir el UCL mediante el percentil de la distribución chi cuadrado cuando $m > p^2 + 3p$.

Las estadísticas $T_{MVE,i}^2$ y $T_{MCD,i}^2$ convergen en distribución a una distribución chi cuadrado con p grados de libertad cuando $m \rightarrow \infty$, pero su distribución es desconocida para valores pequeños o incluso moderados de m . Para estos casos, entonces, los límites de control de las cartas de control T^2 se deben calcular usando simulación de Monte Carlo.

Vargas (2003) compara seis cartas de control ante la presencia de varios *outliers*: las cartas T_u^2 , T_d^2 , T_{MVE}^2 , T_{MCD}^2 y dos cartas adicionales que utilizan estimadores recortados. La carta T_u^2 es la mejor ante la presencia de un solo *outlier*, pero es ineficiente ante la presencia de varios *outliers*. Cuando ocurre más de un *outlier*, las cartas robustas MVE y MCD presentan las probabilidades de señal más altas, siendo el método MVE el mejor. Jensen *et al.* (2006) realizaron un extenso trabajo de simulación en el que compararon los métodos usual, MVE y MCD para diferentes valores de m , p y k (número de *outliers*). Recomiendan el método usual cuando hay a lo más un *outlier*, el método MVE cuando $m \leq 50$ y el método MCD para $m > 50$. Chenouri *et al.* (2009) proponen una carta de control T^2 robusta basada en estimadores de determinante de covarianza mínimo reponderados (RMCD). Los estimadores MCD

reponderados, es decir, calculados en dos etapas, mantienen las propiedades de los estimadores MCD iniciales, afín equivariantes, robustos y normales asintóticamente, mientras que alcanzan una mayor eficiencia. Proponen también usar esta carta en la Fase II de monitoreo.

Yáñez *et al.* (2003, 2010) sugieren utilizar estimadores S para la construcción de la carta T^2 . Los estimadores S fueron propuestos por Rousseeuw & Yohai (1984). El estimador S biponderado de localización y forma se define como el par $(\mathbf{t}, \mathbf{C})$ que minimiza el determinante $|k^2 \mathbf{C}|$, tal que

$$m^{-1} \sum_{i=1}^m \rho \left(\sqrt{(\mathbf{x}_i - \mathbf{t})^t (k^2 \mathbf{C})^{-1} (\mathbf{x}_i - \mathbf{t})} \right) = b_0$$

donde ρ es la función biponderada de Tukey dada por:

$$\rho(d, c) = \begin{cases} \frac{d^2}{2} - \frac{d^4}{2c^2} + \frac{d^6}{6c^4}, & 0 \leq d \leq c \\ \frac{c^2}{6}, & d > c \end{cases}$$

Las constantes c y b_0 se ajustan de tal forma que se logre un punto de ruptura cercano a 0,5.

Yáñez *et al.* (2003, 2010) compararon las cartas usual, MVE y la basada en estimadores S . Nuevamente las simulaciones mostraron que el método usual es el mejor para el caso de un *outlier*. Ante la presencia de varios *outliers*, las cartas de control basadas en los estimadores MVE y S tienen un comportamiento similar.

5.3 Carta de control T^2 basada en estimadores DG

La eficiencia de las cartas de control robustas ante la presencia de varios *outliers* conduce a considerar otros tipos de estimadores dentro de la gran gama de estimadores robustos que se han propuesto en la literatura. En este contexto, se presenta a continuación una propuesta de carta T^2 que usa estimadores basados en funciones de profundidad, conocidos también como estimadores DG.

5.3.1 Estimadores DG (Donoho-Gasko)

Donoho & Gasko (1992) propusieron estimadores multivariados de localización basados en funciones de profundidad. Inicialmente se revisan algunos conceptos generales sobre lo que se conoce como profundidad de un punto.

Nociones generales sobre función de profundidad

El concepto de función de profundidad estadística se ha constituido en una herramienta útil para el análisis de datos multivariados. El problema general consiste en caracterizar o medir qué tan profundo está localizado un punto en una nube de datos. Tukey (1977) fue quien primero presentó el concepto de profundidad para el caso univariado. La profundidad de un valor x en un conjunto unidimensional $X = \{x_1, \dots, x_m\}$ es el mínimo entre el número de puntos que quedan a la izquierda y a la derecha de x :

$$PF_1(x; X) = \min(\#\{i \mid x_i \leq x\}, \#\{i \mid x_i \geq x\}) \quad (5.3)$$

Esta definición se puede extender al caso multivariado. Sea $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ un conjunto de observaciones en un espacio de dimensión p . La profundidad de un punto $\mathbf{x} \in \mathbb{R}^p$ en $\mathbf{X}$ es la mínima profundidad de $\mathbf{x}$ en cualquier proyección unidimensional de $\mathbf{X}$. Es decir, si $\mathbf{u}$ es un vector en $\mathbb{R}^p$ de norma unitaria, entonces el conjunto $\{\mathbf{u}^t \mathbf{x}_i\}$ es una proyección unidimensional de $\mathbf{X}$, y

$$PF_p(\mathbf{x}; \mathbf{X}) = \min_{\|\mathbf{u}\|=1} PF_1(\mathbf{u}^t \mathbf{x}; \{\mathbf{u}^t \mathbf{x}_i\}).$$

El concepto de profundidad permite tener, de alguna manera, un orden en un conjunto de observaciones multivariadas. Se establece que los puntos más profundos en un conjunto de datos deben tener la profundidad numérica más grande. Por ejemplo, bajo la definición en (5.3), la mediana es el valor más profundo en el caso unidimensional. El valor más profundo en un espacio p -dimensional puede, por tanto, pensarse como una mediana multidimensional. Naturalmente, pueden existir muchas formas para determinar algún tipo de orden en una nube de puntos multivariados; por ejemplo, se puede usar la distancia de Mahalanobis para medir profundidad.

Liu & Singh (1997) proponen la profundidad de Mahalanobis de la siguiente forma:

$$PF_M(\mathbf{x}) = [1 + (\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})]^{-1};$$

así, cuanto más cercano esté un punto a la media, en términos de la distancia de Mahalanobis, mayor es la profundidad de Mahalanobis.

Profundidad semiespacial

Sea $\mathbf{x} \in \mathbb{R}^p$. Sea H un semiespacio cerrado que contiene al punto $\mathbf{x}$ y sea $P(H)$ la probabilidad de que una observación ocurra en el semiespacio H . Entonces, la probabilidad semiespacial (HD) de un punto $\mathbf{x}$ en $\mathbb{R}^p$ con respecto a una medida de probabilidad P , es el valor más pequeño de $P(H)$ entre todos los semiespacios H que contiene a $\mathbf{x}$. Es decir,

$$HD(\mathbf{x}; P) = \inf\{P(H) \mid H \text{ es un semiespacio cerrado que contienen a } \mathbf{x}\}.$$

Existen varias aproximaciones para calcular la profundidad semiespacial. Se presentan acá los detalles de la aproximación utilizada por las funciones que se usan en las simulaciones al final del capítulo (ver Wilcox, 2005). Para otras aproximaciones, ver Rousseeuw & Struyf (1998), Struyf & Rousseeuw (2000) y Wilcox (2003, 2005).

El método inicia calculando alguna medida multivariada de localización, por ejemplo $\hat{\boldsymbol{\theta}}$. Aquí se usa el estimador de localización MCD, aunque se puede usar otro estimador, como por ejemplo el MVE. Para cada punto $\mathbf{x}_i$ se traza una recta L que lo conecta con $\hat{\boldsymbol{\theta}}$. Se proyectan (ortogonalmente) todos los puntos sobre L . Esto es, para cada punto $\mathbf{x}_j$, $j = 1, \dots, m$, se traza una perpendicular a L . El punto donde esta perpendicular intercepta a L es la proyección de $\mathbf{x}_j$. Siguiendo la notación de Wilcox (2005), se tiene:

Para $i = 1, \dots, m$, sea

$$\mathbf{u}_i = \mathbf{x}_i - \hat{\boldsymbol{\theta}}$$

y

$$B_i = \mathbf{u}_i^t \mathbf{u}_i = \sum_{k=1}^p u_{ik}^2.$$

La distancia entre $\hat{\theta}$ y la proyección de $\mathbf{x}_j$ sobre la recta que conecta a $\hat{\theta}$ con $\mathbf{x}_i$ es:

$$D_{ij} = \|\mathbf{t}_{ij}\|,$$

donde $\|\mathbf{t}_{ij}\|$ es la norma euclídeana del vector $\mathbf{t}_{ij}$,

$$\mathbf{t}_{ij} = \frac{W_{ij}}{B_i}(u_{i1}, \dots, u_{ip})^t$$

y

$$W_{ij} = \sum_{k=1}^p u_{ik}u_{jk}, \quad j = 1, \dots, m.$$

Sea d_{ij} la profundidad de $\mathbf{x}_j$ cuando se proyectan todos los puntos sobre la recta que conecta a $\mathbf{x}_i$ con $\hat{\theta}$. Es decir, para i, j fijos, la profundidad del valor proyectado de $\mathbf{x}_j$ es:

$$d_{ij} = \min [\#\{D_{ij} \leq D_{ik}\}, \quad \#\{D_{ij} \geq D_{ik}\}],$$

donde $\#\{D_{ij} \leq D_{ik}\}$ indica el número de valores D_{ik} que satisfacen la desigualdad $D_{ij} \leq D_{ik}$. La profundidad de x_j se define como:

$$\min_i d_{ij}.$$

En pocas palabras, la profundidad de un punto nos indica qué tan cerca está este punto respecto del centro, en una nube de puntos multidimensional. Hay varias formas de establecer esta medida. La que se explica en las páginas 181-183 es una de ellas, pero el concepto de esta es el mismo que aparece explicado en el último párrafo de la página 183.

Estimador de localización DG

Basados en el concepto de profundidad, Donoho & Gasko (1992) proponen un estimador multivariado de localización que es afín equivariante. El método consiste en calcular la profundidad para cada uno de los m puntos, eliminar aquellos puntos que no son muy profundos y promediar los puntos que queden. Como se puede observar, este estimador es una extensión multivariada de la media recortada univariada. Por tal razón se denomina media γ recortada de Donoho-Gasko. Puntos con una profundidad menor que γ se eliminan, y se calcula la media de las observaciones restantes.

Carta de control T^2 con estimadores DG

Como se vio en la sección anterior, Donoho & Gasko (1992) propusieron un estimador multivariado de localización mediante el recorte de cierto porcentaje de observaciones, el cual se efectúa de acuerdo con la profundidad de los datos. Luego de eliminar las observaciones menos profundas, el estimador de localización DG es el promedio de las observaciones restantes. En este trabajo se propone usar dichas observaciones restantes para calcular no solo el promedio sino la matriz de varianzas covarianzas usual. De esta manera se obtienen estimadores de localización y dispersión, los cuales se denotan acá como estimadores DG. Las estadísticas T^2 de Hotelling usando estimadores DG se denotan por:

$$T_{DG,i}^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{DG})^t \mathbf{S}_{DG}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{DG}), \quad i = 1, \dots, m.$$

Como se pudo observar arriba, el algoritmo para calcular los estimadores DG es extenso, lo cual hace que los programas que los calculan tomen mucho tiempo. Este es quizás el punto más débil para el cálculo y uso de estos estimadores. En razón a esto, los resultados de las simulaciones que se presentan a continuación son limitados.

El límite de control superior para la carta T^2 se calculó con base en 1,000 replicaciones para el caso $m = 30$, $p = 2$ y $\gamma = 0,1$, de tal forma que la probabilidad global de falsa alarma sea de 0,05. El valor obtenido fue $UCL = 39,6807$.

La carta de control T^2 con estimadores DG se comparó con la carta T^2 usual y la carta de control T^2 basada en estimadores MVE, generando aleatoriamente k outliers para $k = 1, 2, 3$ y estimando la probabilidad de señal a partir de 1,000 replicaciones. Las figuras 5.1 a 5.3 muestran estas probabilidades.

Se puede observar el buen desempeño de la carta de control T^2 con estimadores DG para un número pequeño de outliers. Para un outlier es superada solo por la carta usual. Para dos y tres outliers, las probabilidades son muy similares a la carta basada en estimadores MVE y siempre mejorando la carta usual. Sin embargo, cabe recordar que se está trabajando con estimadores recortados multivariados, y que estas simulaciones se han efectuado eliminando las observaciones que presentaban una profundidad menor que 0,1. Para una investigación futura

queda cambiar el porcentaje de observaciones eliminadas, con el fin de aumentar la potencia de la carta ante la presencia de un número mayor de *outliers*.

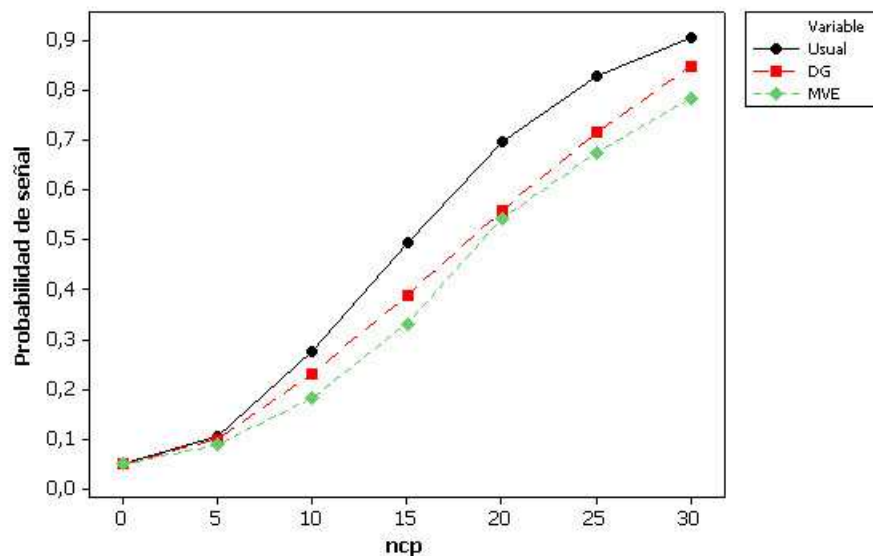


FIGURA 5.1 Probabilidad de señal para un *outlier*. Comparación de los métodos usual, DG y MVE. Donde ncp representa el parámetro de no centralidad.

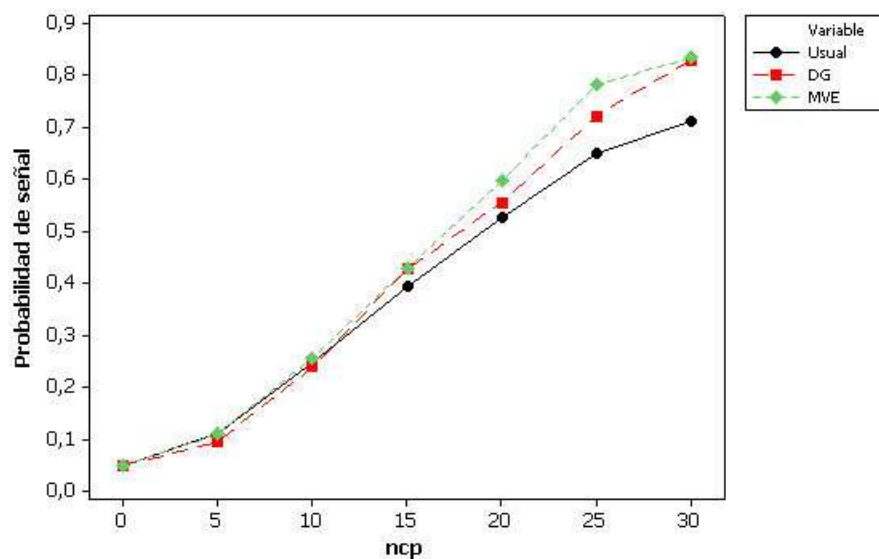


FIGURA 5.2 Probabilidad de señal para dos *outliers*. Comparación de los métodos usual, DG y MVE.

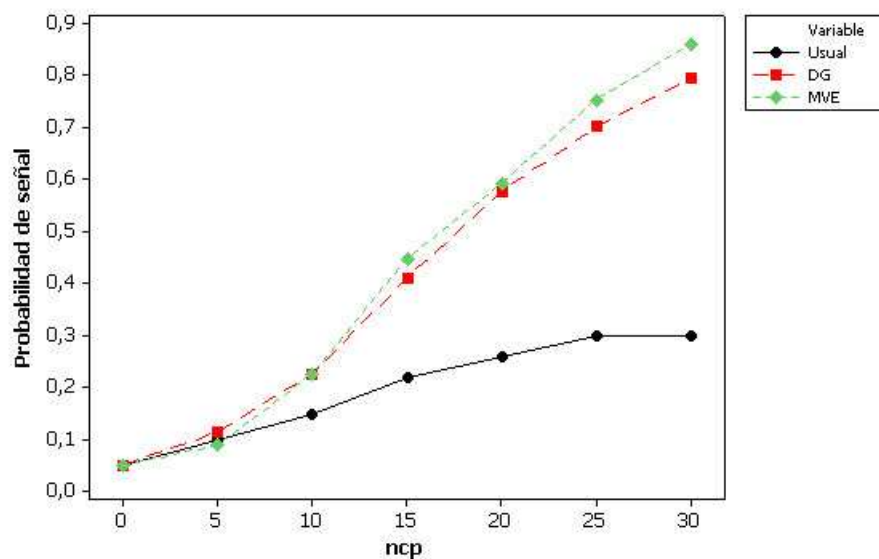
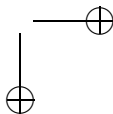
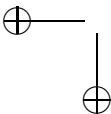
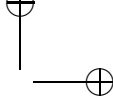
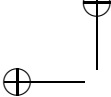


FIGURA 5.3 Probabilidad de señal para tres *outliers*. Comparación de los métodos usual, DG y MVE.



Bibliografía

- [1] Bissell, D. (1994). *Statistical Methods for SPC and TQM*. Londres: Chapman and Hall.
- [2] Campbell, N. A. (1980). “Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation”, *Applied Statistics*, 29, 231-237.
- [3] Chenouri, S., Variyath, A. M. & Steiner, S. H. (2009). “A Multivariate Robust Control Chart for Individual Observations”, *Journal of Quality Technology*, 41(3), 259-271.
- [4] Conover, W. J. (1998). *Practical Nonparametric Statistics* (3th ed.). New York: John Wiley & Sons, Inc.
- [5] Correa, J. C. & González, N. G. (2002). *Gráficos estadísticos con R*. <http://cran.r-project.org/doc/contrib/grafi3.pdf>.
- [6] Crosier, R. (1986). “A New Two - Side Cumulative Sum Quality Control Scheme”, *Technometrics*, 28, 187-194.
- [7] Crosier, R. (1988). “Multivariate Generalizations Of Cumulative Sum Quality - control schemes”, *Technometrics*, 30, 291-303.
- [8] Cryer, J. D. & Ryan, T. P. (1990). “The estimation of sigma for an X chart: $\overline{MR}/d_2$ or S/c_4 ?”, *Journal of Quality Technology*, 22, 187-192.

- [9] Davis, C. M. & Adams, B. M. (2005). “Robust Monitoring of Contaminated Data”, *Journal of Quality Technology*, 37, 163-174.
- [10] De Mast, J. & Roes, K. C. B. (2004). “Robust Individuals Control Chart for Exploratory Analysis”, *Quality Engineering*, **16**(3), 407-421.
- [11] Donoho, D. L. & Gasko, M. (1992). “Breakdown Properties of Location Estimates Based on Half - Space Depth and Projected Outlyingness”, *Annals of Statistics*, 20, 1803-1827.
- [12] Díaz M., L. G. (2002). *Estadística multivariada. Inferencia y métodos*. Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Bogotá.
- [13] Fuchs, C. & Benjamin, Y. (1994). “Multivariate Profile Charts for Statistical Process Control”, *Technometrics*, 36, 182-195.
- [14] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley & Sons, Inc.
- [15] Guerrero G., V. M. (2003). *Análisis estadístico de series de tiempo económicas* (2^a ed.). México: Thomson Learning.
- [16] Harter, H. L. (1960). “Tables of range and studentized range”, *Annals of Mathematical Statistics*, **31**(4), 1122-1147.
- [17] Healy, J. D., (1987). “A note on Multivariate Cusum Procedures”, *Technometrics*, 29, 409-412.
- [18] Holmes, D. S. & Mergen, A. E. (1993). “Improving the Performance of the T^2 Control Chart”, *Quality Engineering*, 5, 619-625.
- [19] Hotelling, H. (1931). “The Generalization of Student’s Ratio”, *Annals of Mathematical Statistics*, 2, 360-378.
- [20] Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley and Sons, Inc.
- [21] Jackson, J. E. (1991). *A User’s Guide to Principal Components*. New York: John Wiley & Sons, Inc. pp. 73.
- [22] Jensen, W. A., Jones-Farmer, L. A., Champ, C. W. & Woodall, W. H. (2006). “Effects of Parameter Estimation on Control Chart

- Properties: A Literature Review”, *Journal of Quality Technology*, **38**(4), 349-364.
- [23] Jensen, W. A., Birch, J. B. & Woodall, W. H. (2007). “High Break-down Estimation Methods for Phase I Multivariate Control Charts”, *Quality and Reliability Engineering International*, **23**(5), 615-629.
- [24] Johnson, R. A. & Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis* (4th ed.), pp. 157-222. New Jersey: Prentice Hall.
- [25] Khoo, M. B. C. (2004). “An Alternative Q Chart Incorporating a Robust Estimator of Scale”, *Journal of Modern Applied Statistical Methods*, **3**, 72-84.
- [26] Kourty, T. & MacGregor, J. F. (1996). “Multivariate SPC Methods for Process and Product Monitoring”, *Journal of Quality Technology*, **28**, 409-428.
- [27] Langenberg, P. & Iglewicz, B. (1986). “Trimmed Mean $\bar{X}$ and R Charts”, *Journal of Quality Technology*, **18**, 152-161.
- [28] Liu, R. G. & Singh, K. (1997). “Notions of Limiting P Values Based on Data Depth and Bootstrap”, *Journal of the American Statistical Association*, **92**, 266-277.
- [29] Lowry, C. A. & Montgomery, D. C. (1995). “A Review of Multivariate Control Charts”, *IIE Transactions*, **27**, 800-810.
- [30] Lowry, C. A., Woodall, W. H., Champ, C. W. & Ridgdon, S. E. (1992). “A multivariate Exponentially Weighted Moving Average Control Chart”, *Technometrics*, **34**, 46-53.
- [31] Mardia, K. V. (1970). “Measures of Multivariate Skewness and Kurtosis with Applications”, *Biometrika*, **57**, 519-530.
- [32] Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis* (pp. 1-86). London: Academic Press.
- [33] Maronna, R. A. (1976). “Robust M - Estimations of Multivariate Location and Scatter”, *Annals of Statistics*, **4**, 51-67.
- [34] Mason, R. L., Chou, Y., Sullivan, J. H., Stoumbos, Z. G. & Young, J. C. (2003) “Systematic Patterns in T^2 Control Charts”, *Journal of Quality Technology*, **35**, 47-58.

-
- [35] Mason, R. L., & Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia: ASA-SIAM.
 - [36] Mason, R. L., Chou, Y. & Young, J. C. (2001). “Applying Hotelling’s T^2 Statistic to Batch Process”, *Journal of Quality Technology*, 33, 466-479.
 - [37] Mason, R. L., Tracy, N. D. & Young, J. C. (1997). “A Practical Approach for Interpreting Multivariate T^2 Control Chart Signals”, *Journal of Quality Technology*, 29(4), 396-406.
 - [38] Mason, R. L., Tracy, N. D. & Young, J. C. (1995). “Decomposition of T^2 for Multivariate Control Chart Interpretation”, *Journal of Quality Technology*, 27, 99-108.
 - [39] Montgomery, D. C. & Peck, E. (1992). *Introduction to Linear Regression Analysis*. New York: Wiley & Sons, Inc.
 - [40] Quesenberry, C. P. (1993). “The effect of sample size on estimated limites for $\bar{X}$ and X control charts”, *Journal of Quality Technology*, 25(4), 237-247.
 - [41] Rencher, A. C. (1995). *Methods of Multivariate Analysis* (pp. 94-120). New York: John Wiley & Sons, Inc.
 - [42] Rocke, D. M. (1989). “Robust Control Charts”, *Technometrics*, 31, 173-184.
 - [43] Rocke, D. M. (1992). “ $\bar{X}_Q$ and R_Q : Robust Control Charts”, *The Statistician*, 41, 97-104.
 - [44] Rousseeuw, P. J. (1984). “Least Median of Squares Regression”, *Journal of American Statistical Association*, 79, 871-880.
 - [45] Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, Inc.
 - [46] Rousseeuw, P. J. & Struyf, A. (1998). “Computing Location Depth and Regression Depth in Higher Dimensions”, *Statistics and Computing*, 8, 193-203.
 - [47] Rousseeuw, P. J. & Van Driessen, K. (1999). “A Fast Algorithm for the Minimum Covariance Determinant Estimator”, *Technometrics*, 41, 212-223.

- [48] Rousseeuw, P. J. & Van Zomeren, B. C. (1990). “Unmasking Multivariate Outliers and Leverage Points (with discussion)”, *Journal of the American Statistical Association*, 85, 633-651.
- [49] Rousseeuw, P. J. & Yohai, V. (1984). “Robust Regression by Means of Estimators”. In: *Robust and Nonlinear Time Series Analysis*, edited by J. Franke, W. Härdle & R. D. Martin, Lectures Notes in Statistics No. 26, Springer Verlag, New York, pp. 256-272.
- [50] Rousseeuw, P. J., Van Alest, S., Van Driessen, K. & Agullo, J. (2004). “Robust Multivariate Regression”, *Technometrics*, 46, 293-305.
- [51] Ryan, T. P. (2000). *Statistical Methods for Quality Improvement* (2nd ed.). New York: Wiley & Sons, Inc.
- [52] SAS/QC, (1999). *SAS/QC, User’s Guide*. SAS Institute Inc.
- [53] Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley & Sons, Inc.
- [54] Struyf, A. & Rousseeuw, P. J. (2000). “High-Dimensional Computation of the Deepest Location”, *Computational Statistics and Data Analysis*, 34, 415-426.
- [55] Sullivan, J. H. & Woodall, W. H. (1996). “A Comparison of Multivariate Control Charts for Individual Observations”, *Journal of Quality Technology*, **28**(4), 398-408.
- [56] Tukey, J. W. (1977). *Exploratory Data Analysis*. Massachusetts: Addison-Wesley, Reading.
- [57] Vargas, J. A. (2001). *Introducción al control estadístico de calidad*. Universidad Nacional de Colombia, Departamento de Matemáticas y Estadística. Bogotá.
- [58] Vargas, J. A. (2003). “Robust Estimation in Multivariate Control Charts for Individual Observations”, *Journal of Quality Technology*, **35**(4), 367-376.
- [59] Williams, J. D., Woodall, W. H., Birch, J. B. & Sullivan, J. H. (2006). “Distribution of Hotelling’s T^2 Statistic Based on the Successive Differences Estimator”, *Journal of Quality Technology*, **38**(3), 217-229.

-
- [60] Wilcox, R. R. (2003). “Approximating Tukey’s Depth”, *Communications in Statistics, Simulations and Computations*, **32**, 977-985.
 - [61] Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing* (2nd ed.). United States: Elsevier Academics Press.
 - [62] Woodall, W. H. (2000). “Controversies and Contradictions in Statistical Process Control”, *Journal of Quality Technology*, **32**(4), 341-350.
 - [63] Woodall, W. H. & Ncube, M. M. (1985). “Multivariate CUSUM Quality Control Procedures”, *Technometrics*, **27**, 285-292.
 - [64] Yáñez, S., Vargas, J. A. & González, N. (2003). “Carta T^2 con base en estimadores robustos de los parámetros”, *Revista Colombiana de Estadística*, **26**(2), 159-179.
 - [65] Yáñez, S., González, N. & Vargas, J. A. (2010). “Hotelling’s T^2 control charts Based on Robust Estimators”, *Dyna*, **77**(163), 239-247.

APÉNDICE A

Una breve introducción a R

R, más que un paquete estadístico de dominio público, es un lenguaje y un ambiente que, además de ofrecer una amplia gama de métodos estadísticos, también puede ser considerado como un lenguaje de alto nivel. Entre otras cosas, permite definir funciones que pasan a ser parte del sistema, las cuales pueden ser usadas en sesiones posteriores. Entre sus ventajas están su capacidad para operar con objetos, la programación en lenguaje matricial, la disponibilidad de una amplia base de operadores y la versatilidad que ofrece en la realización de gráficas.

El software está disponible bajo los términos de licencia GNU en forma de código fuente, para sistemas operativos Windows, Mac OS, UNIX y similares, y puede ser descargado en el website www.r-project.org, donde se encuentran disponibles referencias importantes para aprender su uso, tales como:

1. *An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics, Version 2.13.0 (2011-04-13)*, by *W. N. Venables, D. M. Smith and the R Development Core Team*;

2. *R: A Language and Environment for Statistical Computing Reference Index, Version 2.13.0 (2011-04-13), by The R Development Core Team.*

Al iniciar una sesión en R, el programa se ejecuta y abre la ventana de comandos denominada R-Console, en la cual se pueden escribir comandos que serán ejecutados una vez se presione Enter. Como alternativa, se puede recurrir a la creación de un archivo de edición con el menú Archivo - Nuevo Script, para editar los programas que se desee y ejecutarlos en forma completa o parcialmente mediante selección previa, con las opciones disponibles en el menú Editar. Estos archivos de edición se pueden guardar con extensión .R y ser utilizados en cualquier otra sesión cargándolos por el menú Archivo - Abrir Script.

En R, las funciones están organizadas en librerías o paquetes. Por defecto, R inicializa en el paquete denominado **base**, en el cual se concentran las funciones generales para el manejo de datos y gráficas. Existen otros paquetes en los cuales se encuentran herramientas de análisis más especializadas; estas pueden ser utilizadas cargando previamente la librería que las contiene. Una librería o paquete se puede cargar mediante la función `library()` o bien `require()` o a través del menú Paquetes del R. Una descripción de estas librerías se puede obtener ejecutando:

```
> library()
```

lo que permite ver en una ventana un listado como el siguiente:

```
Packages in library 'C:/PROGRA~1/R/R-212~1.2/library':
```

akima	Interpolation of irregularly spaced data
base	The R Base Package
boot	Bootstrap R (S-Plus) Functions (Canty)
car	Companion to Applied Regression
class	Functions for Classification
cluster	Cluster Analysis Extended Rousseeuw et al.
codetools	Code Analysis Tools for R
datasets	The R Datasets Package
forecast	Forecasting functions for time series
foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...
fracdiff	Fractionally differenced ARIMA aka ARFIMA(p,d,q) models
graphics	The R Graphics Package
grDevices	The R Graphics Devices and Support for Colours

	and Fonts
grid	The Grid Graphics Package
KernSmooth	Functions for kernel smoothing for Wand & Jones (1995)
lattice	Lattice Graphics
leaps	regression subset selection
locfit	Local Regression, Likelihood and Density Estimation.
lpSolve	Interface to Lp_solve v. 5.5 to solve linear/integer programs
MASS	Support Functions and Datasets for Venables and Ripley's MASS
Matrix	Sparse and Dense Matrix Classes and Methods
methods	Formal Methods and Classes
mgcv	GAMS with GCV/AIC/REML smoothness estimation and GAMMs by PQL
nlme	Linear and Nonlinear Mixed Effects Models
nnet	Feed-forward Neural Networks and Multinomial Log-Linear Models
quadprog	Functions to solve Quadratic Programming Problems.
rpart	Recursive Partitioning
sampling	Survey Sampling
spatial	Functions for Kriging and Point Pattern Analysis
splines	Regression Spline Functions and Classes
stats	The R Stats Package
stats4	Statistical Functions using S4 Classes
survey	analysis of complex survey samples
survival	Survival analysis, including penalised likelihood.
tcltk	Tcl/Tk Interface
tools	Tools for Package Development
TSA	Time Series Analysis
tseries	Time series analysis and computational finance
utils	The R Utils Package
zoo	Z's ordered observations

Existe una gran cantidad de paquetes disponibles en el sitio web de R que el usuario puede descargar simplemente mediante los siguientes pasos:

1. Conéctese a internet.
2. Inicialice una sesión R. Vaya al menú Paquetes-Seleccionar Espejo CRAN.
3. En la ventana CRAN mirror seleccione Colombia y presione OK.
4. Vuelva al menú Paquetes y esta vez seleccione Instalar paquete(s), y en la ventana Packages resultante seleccione el paquete que se desea

descargar, buscando por nombre en orden alfabético, y finalmente presione OK.

Para consultar la lista de funciones disponibles en una librería particular se usa el comando `library(help="nombre de la libreria")`; por ejemplo, si se desea saber cuáles funciones están disponibles en la librería `splines`, se ejecuta el siguiente comando:
`library(help="splines")`.

Para conocer sobre la sintaxis y el uso de alguna función en particular, se puede usar el menú Ayuda o el comando `?paquete::funcion`, por ejemplo, `?car::boxCox`.

Nota. R es sensible a mayúsculas y minúsculas, por lo que es importante escribir los nombres de funciones y librerías conforme han sido definidos.

Algunas veces el usuario desconoce el nombre de la función deseada. Si ya existe en alguna librería previamente descargada, se puede recurrir a la búsqueda con una palabra clave con la función `help.search("palabra-clave")` o `??palabra-clave`. Por ejemplo, `??shapiro` o `help.search("shapiro")` da como resultado un listado de funciones R y su ubicación (`paquete::funcion`), como el siguiente:

```
Help files with alias or concept or title matching 'shapiro' using
fuzzy matching:
stats::shapiro.test      Shapiro-Wilk Normality Test

Type '?PKG::F00' to inspect entries 'PKG::F00', or 'TYPE?PKG::F00' for
entries like 'PKG::F00-TYPE'.
```

Para la lectura de bases de datos externas R cuenta con las funciones `scan()`, `read.table()`, `read.csv()`, `read.delim()`, `read.fwf()`.

Los resultados numéricos de alguna función o comando se pueden explorar con las opciones Copiar - Pegar o enviándolos a archivos externos con las funciones R `sink()`, `write()`, `dump()`; esta última guarda objetos en formato R (matrices, vectores, arreglos, etc.). De igual manera, las gráficas producidas se pueden guardar en una variedad de formatos (metafile, postscript, png, pdf, bmp, tiff, jpeg), o copiarse y pegarse directamente en un editor de texto como MS-Word. Datos y comandos guardados como objetos R (con función `dump()`) se pueden cargar en cualquier sesión con la función `source()`.

R además permite al usuario crear sus propias funciones con el comando `function()`; estas son almacenadas en una forma interna especial y pueden ser usadas en expresiones subsiguientes. Esta característica proporciona una gran potencialidad y versatilidad al lenguaje, permitiendo al usuario mayor productividad en el uso de R. Una función es definida mediante una asignación de la forma:

```
nombre=function(arg.1, arg.2, \ldots){expresiones}
```

Las expresiones usadas son expresiones admisibles en R que usan los argumentos `arg.i`, para calcular un valor o producir un objeto determinado (vector, matriz, lista, arreglo, etc.). Una invocación o llamada de la función toma la forma `nombre(expr.1,expr.2,...)` donde `expr.i` es el valor particular para el argumento `arg.i` de la función.

A continuación, un listado de las funciones básicas utilizadas en este libro o que pueden ser útiles para el lector. Para más detalles, se sugiere consultar directamente la ayuda y los manuales de R.

Funciones para entrada y lectura de datos y escritura y exportación de resultados y objetos R

Comando	Función
<code>scan()</code>	Lectura de datos. Especial para datos no estructurados
<code>read.table()</code>	Lectura de archivos en formato de tabla
<code>read.fwf()</code>	Lectura de archivos en formato de tabla con ancho fijo
<code>read.csv()</code>	Lectura de archivos en formato de tabla con datos separados por comas
<code>read.ftable()</code>	Lectura de una tabla de contingencia guardada en archivo ASCII
<code>sink()</code>	Envío a archivo ASCII los resultados de una sesión
<code>write()</code>	Escribe una matriz en un archivo ASCII
<code>write.ftable()</code>	Escribe una tabla de contingencia en un archivo ASCII
<code>xtable()</code>	Escribe una matriz en formato Latex. Requiere la librería <code>xtable</code>
<code>ftable()</code>	Permite presentar decentemente un arreglo multidimensional

Operadores

Aritméticos	De comparación	Lógicos y de control
+ Suma	< menor	& y
- Resta	> mayor	ó
* Multiplicación	<= menor o igual	! no
/ División	>= mayor o igual	all(...) ¿Todos los valores lógicos son ciertos?
^ Exponenciación	== igual	any(...) ¿Alguno de los valores lógicos es cierto?
%/% División entera	!= diferente	&& Si primer operando es cierto evalúa segundo operando
%% Operador módulo		Si primer operando es falso evalúa segundo operando.

Funciones relacionadas con distribuciones

Distribución	Densidad	Función acumulada	Cuantil p	Números aleatorios
Uniforme	dunif(x,...)	punif(q,...)	qunif(p,...)	runif(n,...)
Normal	dnorm(x,...)	pnorm(q,...)	qnorm(p,...)	rnorm(n,...)
Binomial	dbinom(x,...)	pbinom(q,...)	qbinom(p,...)	rbinom(n,...)
Lognormal	dlnorm(x,...)	plnorm(q,...)	qlnorm(p,...)	rlnorm(n,...)
Beta	dbeta(x,...)	pbeta(q,...)	qbeta(p,...)	rbeta(n,...)
Geométrica	dgeom(x,...)	pgeom(q,...)	qgeom(p,...)	rgeom(n,...)
Gamma	dgamma(x,...)	pgamma(q,...)	qgamma(p,...)	rgamma(n,...)
Ji cuadrado	dchisq(x,...)	pchisq(q,...)	qchisq(p,...)	rchisq(n,...)
Exponencial	dexp(x,...)	pexp(q,...)	qexp(p,...)	rexp(n,...)
F	df(x,...)	pf(q,...)	qf(p,...)	rf(n,...)
Hipergeom.	dhyper(x,...)	phyper(q,...)	qhyper(p,...)	rhyper(n,...)
t	dt(x,...)	pt(q,...)	qt(p,...)	rt(n,...)
Poisson	dpois(x,...)	ppois(q,...)	qpois(p,...)	rpois(n,...)
Weibull	dweibull(x,...)	pweibull(q,...)	qweibull(p,...)	rweibull(n,...)
Binom. Neg.	dnbinom(x,...)	pnbinom(q, s,...)	qnbinom(p,...)	rnbinom(n,...)

Condicionales y loops

- if(condición){expresiones} else expresión
- ifelse(condición,1,0)
- for(nombre in expresión){expresiones}
- while(condición){expresiones}

Funciones que producen escalares

Comando	Función
max()	Máximo del argumento
min()	Mínimo del argumento
sum()	Suma de todos los elementos del argumento
mean()	Promedio aritmético de todos los elementos del argumento
var()	Varianza de todos los elementos del argumento, cuando este es un vector, o matriz de covarianzas si el argumento es una matriz
median()	Mediana del argumento
quantile(...,probs=c(...))	Cuantiles del argumento con las proporciones indicadas en 'probs'
prod()	Producto de todos los elementos del argumento
length()	Número de elementos del argumento si este es una lista o vector
ncol()	Número de columnas si el argumento es una matriz
nrow()	Número de filas si el argumento es una matriz

Algunos objetos R

Vectores, matrices y arreglos	
Comando	Función
c()	Crea vectores
append()	Combina vectores o adiciona elementos a un vector
matrix(), as.matrix(), data.matrix()	Crea matrices
array(), as.array()	Crea arreglos
Listas y tramas de datos	
Comando	Función
list(), as.list()	Crea listas de objetos
data.frame(), as.data.frame()	Crea colecciones de variables en estructura tabular

Algunas operaciones con matrices

Comando	Función
%*%	Producto matricial
t()	Transposición de una matriz
crossprod(A)	Producto $A^t A$
svd()	Descomposición en valores singulares de una matriz
qr()	Descomposición qr
chol()	Descomposición de Cholesky
solve()	Inversa de una matriz
cbind()	Combina matrices por columnas
rbind()	Combina matrices por filas
eigen()	Cálculo de valores y vectores propios
diag()	Crea una matriz diagonal si el argumento es un vector o retorna la matriz diagonal de una matriz

Aplicando funciones a objetos R

Comando	Función
<code>apply()</code>	Aplica una función a filas o columnas de una matriz
<code>tapply()</code>	Aplica una función a cada celda de un arreglo
<code>lapply()</code>	Aplica una función a cada elemento de una lista y devuelve una lista
<code>sapply()</code>	Como <code>lapply</code> pero devuelve un vector o matriz

Funciones que producen gráficas

Comando	Función
<code>hist()</code>	Gráfica histogramas
<code>boxplot()</code>	Gráfica boxplots
<code>plot()</code>	Función genérica para gráficos de dispersión, de series de tiempo, de residuales, etc.
<code>qqplot()</code>	Gráfico cuantil-cuantil
<code>qqnorm()</code>	Gráfico de probabilidad normal
<code>pairs()</code>	Gráfico de matrices de dispersión

Funciones varias

Comando	Función
<code>lm()</code>	Función para ajuste de un modelo lineal por mínimos cuadrados
<code>summary()</code>	Función genérica para exhibir resumen de resultados de modelos ajustados
<code>cor.test()</code>	Para probar asociación entre pares de muestras
<code>density()</code>	Función para estimación kernel de una densidad
<code>mahalanobis()</code>	Función que devuelve la distancia de Mahalanobis de las filas de un objeto matricial
<code>mvrnorm()</code>	Función de la librería MASS que produce muestras de una distribución normal multivariada especificada
<code>prcomp()</code>	Función para realizar análisis de componentes principales sobre una matriz de datos
<code>ts()</code>	Crea objetos tipo series de tiempo
<code>acf()</code>	Estima y grafica las autocorrelaciones o autocovarianzas para un objeto serie de tiempo
<code>pacf()</code>	Estima y grafica las autocorrelaciones parciales para un objeto serie de tiempo
<code>Box.test()</code>	Realiza test Box-Pierce o test Ljung-Box sobre un objeto serie de tiempo
<code>lag.plot()</code>	Grafica una serie <i>vs.</i> versiones rezagadas de ella misma

**Cartas de control T^2 multivariadas
usando R y SAS**

Se imprimieron 300 ejemplares en diciembre de 2011 en la Editorial Universidad Nacional de Colombia. En su composición se utilizaron los siguientes elementos: fuente serif romana 11 puntos, formato 16,5 x 24 cm, papel propalcote de 240 g para su carátula y bond de 75 g para las páginas interiores.

